IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 May 2024, accepted 11 May 2024, date of publication 15 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401240

RESEARCH ARTICLE

A Novel Colorectal Histopathological Image Classification Method Based on Progressive Multi-Granularity Feature Fusion of Patch

ZHENGGUANG CAO[1], **WEI JIA**[1,2], **HAIFENG JIANG**[3], **XUEFEN ZHAO**[1], **HONGJUAN GAO**[1,2], **JIALONG SI**[1], **AND CHUNHUI SHI**[1]

[1]School of Information Engineering, Ningxia University, Yinchuan 750021, China
[2]Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West, Yinchuan 750021, China
[3]Department of Pathology, General Hospital, Ningxia Medical University, Yinchuan 750021, China

Corresponding authors: Wei Jia (jiawnx@163.com) and Haifeng Jiang (jhf0347@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62062057 and Grant 12062021, and in part by the Natural Science Foundation of Ningxia Province under Grant 2022AAC03005.

ABSTRACT Colorectal cancer (CRC) is a significant global health concern, ranking as the second most common cancer worldwide. Accurate classification of CRC is crucial for clinical practice and research. Deep learning-based methods have gained popularity in computer-aided CRC classification tasks. However, existing methods often overlook discriminative features at different local granularities, leading to suboptimal classification results. In this paper, we propose a novel Colorectal Histopathological Image Classification Method Based on Progressive Multi-granularity Feature Fusion of Patch (PMFF). Our method combines global features of CRC with features at different local granularities, enhancing the classification process. PMFF employs a progressive learning strategy to guide the model's attention towards information with locally different patch granularity at different stages, culminating in feature fusion at the final stage. The classification method encompasses an information communication mechanism between patches, a feature enhancement strategy, and a feature extraction network for the progressive learning strategy. We conducted evaluations on three public datasets, and the experimental results demonstrate that our method outperforms existing CRC classification methods, achieving classification accuracies of 96.6% and 92.3%, Precisions of 96.5% and 92.4%, Recalls of 96.3% and 92.3%, as well as F1-scores of 96.4% and 92.3%, respectively.

INDEX TERMS Colorectal cancer, progressive learning, multi-granularity, feature extraction network.

I. INTRODUCTION

Cancer has always been a leading cause of increasing mortality worldwide. Colorectal cancer (CRC) has emerged as the second most prevalent cancer globally, posing a significant threat to human health. In terms of incidence, it ranks third. It is projected that the number of CRC cases will reach 3.2 million by 2040, presenting a major global public health challenge [1]. Early screening, prevention, and treatment can effectively reduce the fatality rate of CRC and enhance patient survival. Therefore, there is an urgent need for precise classification techniques to develop accurate treatment plans during the diagnostic process, ultimately improving patient survival outcomes. Different types of CRC vary in etiology, pathology, clinical presentation, and treatment. Utilizing computer-assisted techniques to analyze histopathological images and achieve accurate classification is crucial for both clinical practice and CRC research.

Computers rely on deep learning techniques for pathology image classification. In recent years, in the field of deep

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

learning, Hong et al. proposed a high-resolution domain adaptation network [2] for the cross-city semantic segmentation task, as well as SpectralGPT [3] and Multimodal artificial intelligence foundation models [4] for processing spectral remote sensing data. All the above works have promoted the application of deep learning techniques in the field of images.

For CRC, in the early conventional classification methods, several scholars have proposed traditional classification methods, such as stain normalization, stain enhancement, and domain adaptive or domain generalization techniques, for cancer cell classification and detection tasks [5], [6]. Ciompi et al. [7] emphasized the importance of staining normalization in colorectal tissue classification tasks. They successfully mitigated staining differences between different images through normalization, thereby enhancing the accuracy and reliability of the network in colorectal tissue classification. However, these methods do not focus on the feature extraction work of CRC, which makes it challenging for pathologists to interpret the lesion's status. To address this issue, Wu et al. [8] proposed a model that combines deep and manual features to improve the prediction of the mutation status of colorectal tissues. It is worth noting that the manual feature extraction method used in the classification of pathology images is expensive and prone to errors.

To overcome the limitations of traditional classification methods, some scholars have explored the use of deep learning to address challenges such as color enhancement and manual feature extraction. Recent studies have demonstrated the effectiveness of deep learning, particularly neural networks, in addressing medical image problems including diagnosis, segmentation, and classification prediction [9], [10]. For instance, Ohata et al. [11] applied transfer learning from convolutional neural network architectures, extracting features from images and utilizing support vector machines (SVMs) for classification. Their method achieved strong performance and provided interpretability for colorectal cancer (CRC) prediction results within the field of deep learning. Rachapudi et al. [12] discussed the application of "improved convolutional neural networks (CNN-5B)" for histopathological image classification. Vuong et al. [13] proposed a novel self-supervised contrast learning framework named IMPaSh, which leverages the ResNet50 encoder to extract domain-shift resistant image representations and employs other domain generalization techniques for classifying colorectal tissue images across domains. Kather et al. [14] utilized a combination of feature extraction techniques and SVM (FE-SVM) algorithms to analyze pathological section images of colorectal cancer patients, aiming to identify specific features and correlate them with the patients' survival status. Furthermore, deep learning methods, such as image generation techniques [15], [16], have been widely applied to the classification of CRC. In recent works, Nergiz [17] converted a new generalized visual representation learning method the Big Transfer model, and six classical deep learning methods into a federated version. The proposed model was tested for single learning, centralized learning, and federated learning and achieved good classification performance. Yu et al. [18] described an innovative discriminative manifold distribution alignment(DMDA) method specifically designed to improve medical image diagnosis of colorectal cancer. DMDA goes beyond traditional methods of data analysis, focusing on local and global distribution alignment and learning the inherent geometric features that exist in manifold spaces through complex learning.

Despite the significant advantages of deep learning methods over those using staining enhancement and handcrafted features, they still have some limitations. The aforementioned methods primarily rely on the effectiveness of deep learning in processing image features, which can cause the network to prioritize salient discriminative features of a global nature, such as cell clusters of CRCs, while disregarding discriminative local features of varying sizes, such as scattered nuclei and mesenchyme. It is important to note that the interior of CRCs consists of complex components, including nuclei, mesenchyme, and secretion, and these local features play a crucial role in the classification of CRCs.

As shown in Fig. 1 the CRC image, which includes five common types of CRC labeled as (a)-(e). The green rectangular region in the figure represents features with larger local granularity, while the blue rectangular region represents features with smaller local granularity. Despite the variations in the sizes of each CRC category, they still exhibit two distinct levels of granularity. In the adipose tissue example shown in Fig. 1(a), the green region contains round and oval regions surrounded by mesenchyme, while the blue region contains the nucleus and mesenchyme. In the mucus tissue example in Fig. 1(b), there is a greater accumulation of secretion within the green region and a relatively smaller amount of mesenchyme within the blue region. In the muscle tissue example in Fig. 1(c), the green region has a larger area occupied by the mesenchyme and cell population compared to the blue region. In the case of the stomach tumor tissue shown in Fig. 1(d), the green area represents a closed area formed by cup cells surrounded by glandular cells, while the blue area contains scattered cells. Finally, in the colorectal tumor tissue example in Fig. 1(e), the tumor tissue within the green area is larger, while the accumulation in the blue area is relatively smaller. During the learning process, if only global features are extracted, both local granularity levels will be considered. However, if the learning process is divided into different granularity levels and features are learned separately, it will facilitate the detailed extraction of features. In addition to the above features, CRC also exhibits unique

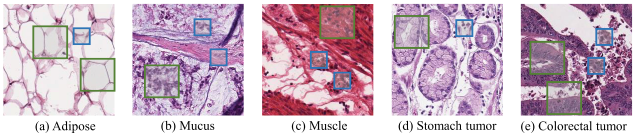

(a) Adipose (b) Mucus (c) Muscle (d) Stomach tumor (e) Colorectal tumor

FIGURE 1. Characteristics and morphological features of CRC at two grain sizes.

morphological characteristics. Notably, the mesenchyme often compresses and distorts the cells, as depicted by the line-like distribution of mesenchyme in Fig. 1(b) and the deformation of the nucleus by the mesenchyme in Fig. 1(c). Overall, CRC displays a relatively intricate organizational structure. To comprehend this complexity, the model must not only extract global features but also learn the detailed features indicated in the figure at different levels of granularity, based on the morphological attributes of CRC.

Existing methods only learn CRC in a single pass, and can only learn global features with discriminative properties, which makes it difficult to cope with the above-proposed feature forms. To address the limitations of current classification methods and based on the features of CRC pathological images, we propose a novel colorectal histopathological image classification method called Progressive Multi-granularity Feature Fusion of Patch (PMFF). The flowchart of PMFF is shown in Fig. 2, which utilizes the unique features of CRC pathological images and employs a progressive learning strategy that consists of three phases. By processing patches with different granularity, we can extract global features and effectively fuse the information from various local features.

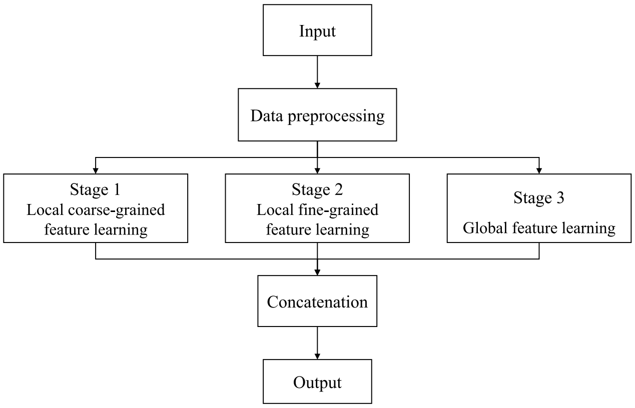


FIGURE 2. **Flowchart of the proposed approach.**

The main contributions of this paper can be summarized as follows:

- We propose a new CRC classification method that categorizes input features into three categories: local coarse-grained feature learning, local fine-grained feature learning, and global feature learning. We utilize a progressive learning strategy to guide the network in learning the original image as a whole and two different local granularities. Finally, we fuse these features.
- We propose information communication mechanisms for local coarse-grained and local fine-grained Patch. These mechanisms are divided into two modules: Channel Interaction and Shifting Module (CISM) and Group Weighted Fusion Module (GWFM). Additionally, we design the Stage-adaptive Feature Enhancement Module (SAFEM) to enhance the original image and the features under two local granularities. This enhancement helps the model locate important information more efficiently.
- We developed a feature extraction network called Progressive Multi-Feature Extraction Network (PMEN) to learn the local features within each patch accurately based on the morphological characteristics of CRC. This network avoids mixing up patch information and effectively extracts multi-granular features of CRC.

II. RELATED WORK

A. PROGRESSIVE LEARNING

Progressive learning is a method that learns information at different granularities in an image step by step. Rather than learning from all granularities at once, this method enables the network to first uncover the overall structure of the image distribution and then focus on smaller granularities. This progressive learning process can ultimately assist the model in capturing more detailed features. Initially introduced in the field of generative adversarial networks [19], progressive learning has gained widespread adoption [20], [21].

In recent years, researchers have begun to apply progressive training methods to image classification tasks and have observed that this method exhibits superior performance in this area. In 2020, Du et al. [22] pioneered a multi-granularity-based progressive training method that attempts to improve fine-grained classification performance using jigsaw data augmentation. The method divides the training process into multiple stages, and at each stage, the model learns feature representations with different granularities and fuses the multi-granularity features in the final stage. The experimental results show that the progressive training strategy can significantly improve the fine-grained classification performance and the visualization results demonstrate that progressive learning can better capture small and complex detailed features in images. Subsequently, Zhang et al. [23] in 2021 proposed a stepwise method of learning to pay attention, which allows the model to gradually focus on fine-grained information. This method allows the model to first focus on global features and then gradually increase the attention to detail to finally improve performance by fusing multi-granularity features. In 2023, Cao et al. [24] applied a progressive learning strategy to the task of classifying non-small cell lung cancer by successively learning vesicular structure features at different granularities, followed by feature fusion. Experiments and visualization images demonstrate that the fusion of features at multi-granularity through progressive learning can also learn features in pathology images well.

The above work fully validates the effectiveness of fusing multi-granularity features using a progressive learning strategy, which is better able to learn complex features. Inspired by the above work, in this work, we draw on the above idea of progressive learning to design a single network that can learn information at different granularities through a series of training stages.

B. SPATIAL FEATURE RECONSTRUCTION

Spatial feature reconstruction is an effective method for training networks that divide an image into multiple local regions and then recombine them, which can result in new feature forms and feature representations of different granularity. The method of dividing or recombining images has been widely used in several domains, including image enhancement [25], weakly supervised tasks [26], and jigsaw puzzles [27]. Not to be overlooked, this method plays an important role in image classification tasks as well, and its main advantage lies in its ability to fully exploit and utilize the spatial relationship information in the image during the training process.

In previous research, methods for segmenting an image into several identical segments have been widely used in image classification tasks. A typical example is a method called ''Tokens-to-Token ViT'' proposed by Yuan et al. [28] in 2021, which allows Vision Transformer (ViT) models to be trained from scratch on large-scale image datasets such as ImageNet. The key idea of the method is to divide pixel-level images into a set of small image chunks, and then convert these image chunks into tokens, similar to vocabulary in natural language processing. These tokens are encoded through an embedding layer and then fed into a Transformer model for processing, and the authors validated the effectiveness of this method through experiments on ImageNet. It is also possible to go a step further and recombine the segmented image chunks into new features for better transformer performance. In 2022, Ren et al. [29] describe how to use a new positional encoding scheme, the Masked Jigsaw Puzzle (MJP), in the Visual Transformer. MJP simply divides a pair of images into grids and randomly disrupts the grid's order, so that each position corresponds to a random position. Experiments have shown that MJP can achieve better performance in a variety of visual tasks by this method.

The above work has shown that spatial feature reconstruction methods can be effective in facilitating and improving performance in classification tasks. However, these methods segment the image into regions of the same size throughout the training process, which means that it is difficult to utilize multi-granularity regions. In this work, we instead form features at two granularities with the above idea of spatial feature reconstruction and work with progressive learning to limit the granularity of the regions learned at each stage.

III. METHOD

To improve the classification accuracy, a novel colorectal histopathological image classification method called PMFF is proposed. As shown in Fig. 3, the proposed method utilizes a progressive learning strategy that divides the learning process into three stages. In these stages, the input original image F is sequentially passed through stage 1, stage 2, and stage 3. The main objective of the first two stages is to extract information from different local granularities. Stage 3 focuses on training the feature extraction ability of the network at the granularity level of the original image and fusing the feature information from different local granularities.

In the first stage, our method focuses on extracting local large-grained information. The original image F is initially processed using a Patch-wise Spatial Restructuring Strategy (PSRS) to combine the spatial features into patches. These patches are then transformed into a linear representation through patch embedding. The CISM and SAFEM are subsequently employed for information transfer and feature enhancement, respectively. The features extracted by PMEN are finally fed into Classifier 1 for classification.

The ability of the PMEN to extract finer features is trained in stage 2. Since the extraction ability of PMEN for local large granularity features is trained in stage 1, it provides a suitable starting point for stage 2. In stage 2, PMEN focuses on smaller granularity based on the previous training. After F is segmented by PSRS, local features with smaller granularity are formed using Patch-wise Feature Matching Strategy (PFMS) and patch embedding. Our PMEN then adaptively mines discriminative information from smaller local details due to the limited receptive domain and representational capability of small regions. The features are successively passed through the GWFM and SAFEM for information transfer and feature enhancement, respectively, before being passed into PMEN. After learning, the features are passed into Classifier 2 for classification.

In the third stage, the main focus is to learn the global features of the CRC. The network has already been able to extract information at different local granularities through learning and parameter updating in the first two stages. Initially, the original image undergoes a linear mapping operation called patch embedding, followed by augmentation of the global features using SAFEM. PMEN can preserve the spatial structure features of CRC while extracting the global features from the original image. After learning, the two types of information with different local granularities learned in the first two stages are combined in the original image to adaptively fuse the features at different granularities.

To guide the network in extracting diverse features at different granularities, classification loss is added at all three stages. The prediction results are computed by Classifier L, the classification layer. For each stage output, the loss is computed using the true label y and the prediction results. The cross-entropy loss function is used for computing the loss, which introduces smoothing noise on the labels of the training data through label smoothing. This reduces overfitting and improves model generalization. Label smoothing is achieved by adjusting the One-Hot coding of the true label y from 1 and 0 to values slightly less than 1 and slightly greater than 0, respectively.

$$L_{CE}(y^L,\ y) = -\sum\nolimits_{i=1}^{K} y_i \log y_i^L \tag{1}$$

where $L = \{1,2,3\}$ denotes the three stages to which it belongs. The loss function for the first stage is L_{CE}^1, which is

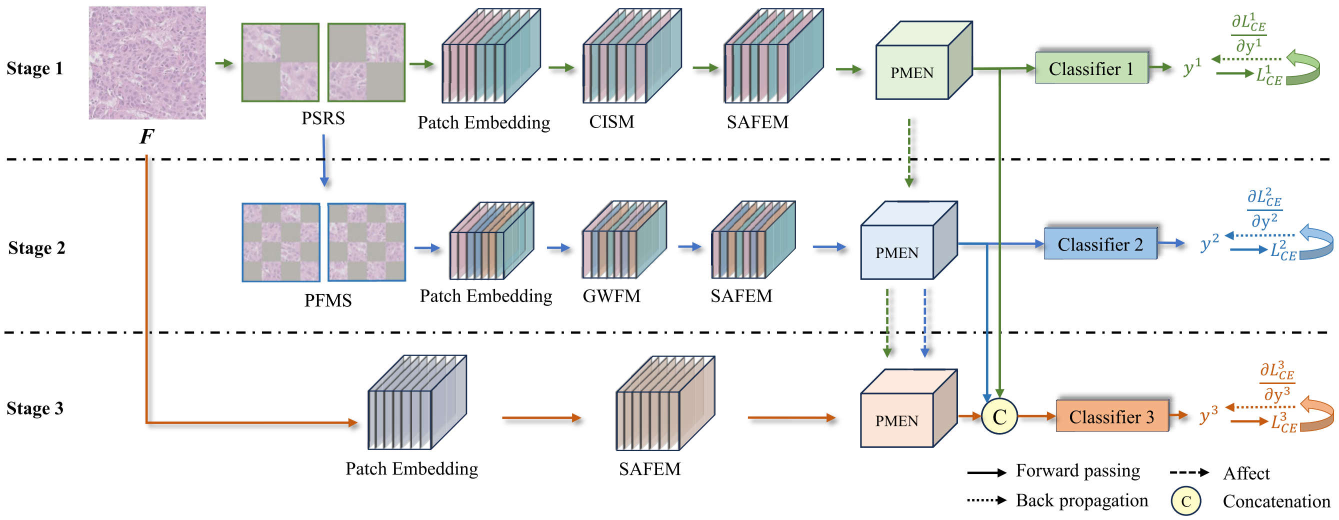


FIGURE 3. Overall structure of the proposed method.

calculated from the prediction result y^1 with the real label y in the first stage and represents the classification loss of large granularity information. The loss function of the second stage is L^2_{CE}, which is calculated from the prediction result y^2 with the real label y and represents the loss of small granularity information. The loss function in the third stage is L^3_{CE}, which is calculated from the prediction result y^3 with the true label y, and represents the loss obtained by classifying the information in the fusion of the global features and the two local granularities of size. Subsequently, the partial derivatives of the loss function and the prediction results are computed at each stage to update the parameters, and in this way, the network can be trained to extract features of different granularity.

A. PATCH-WISE SPATIAL RESTRUCTURING STRATEGY

The features of CRC are typically described in terms of two different local granularities, with variations in the relative sizes presented by each category. For instance, in mucus, the accumulation demonstrates a larger granularity compared to the mesenchyme, while in the tumor, the tumor tissue exhibits a larger granularity relative to the accumulation. Despite these differences, they still share the characteristics of different levels of granularity. The key step that makes the progressive learning strategy effective in capturing the local features of CRCs with different granularity levels is the reconstruction of spatial features in the images. The designed PSRS models the input feature map as a patch with a large granularity, and the first stage in progressive learning focuses on learning the features at that granularity.

The workflow of PSRS is illustrated in Fig. 4. The main task of this module is to slice image F into multiple patches based on their spatial locations and then recombine them. The objective is to effectively remodel the spatial locations of the image. It is important to note that if the size of the first segmented patches is too small, it will result in a significant loss of information in each patch. To address this issue, we choose to divide it equally into four larger patches, namely ‘a’, ‘b’, ‘c’, and ‘d’.

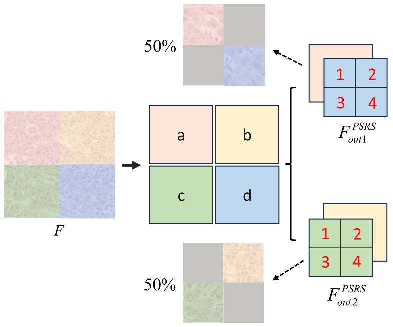


FIGURE 4. PSRS workflow (shaded areas indicate non-existent areas).

In order to avoid slowing down training and having incomplete information, we take into account the variability between features. Our design is partly inspired by the Graph Convolutional Network (GCN) [30] in the field of image processing, which represents image pixels as nodes in a graph. In this method, nodes pay more attention to the information from their neighboring nodes during information transfer, while the more distant information is relatively less important. This confirms that pixels within a neighborhood have stronger correlations with each other compared to diagonal neighborhoods. We have borrowed and applied this idea in our design by combining patches with large differences in features. This allows the network to more easily distinguish and learn more specific features. Specifically, we combine the features of ‘a’ and ‘d’ into F^{PSRS}_{out1}, and the features of ‘b’ and ‘c’ into F^{PSRS}_{out2}, which enables the network to make better use of the feature differences between the patches to enhance the recognition of different local regions.

Considering that PSRS directly splits the space into 4 patches will destroy the original continuous features, and the cut-off feature part will not only lead to incomplete information in the current patch but also become interfering information in other patches. To avoid this situation, we add a trainable parameter α to control the size of the spatial partition. H_P and W_P are the height and width of the large patch, and they satisfy the following equation 2:

$$\begin{cases} H_P = f(H,\ \alpha) = \alpha \times H \\ W_P = f(W,\ \alpha) = \alpha \times W \end{cases} \tag{2}$$

Considering that the complexity of the network cannot be increased after spatial grouping, we consider $\alpha = 0.5$ as the original region of patch, which can be regarded as the minimum criterion for spatial grouping. the part of $\alpha > 0.5$ will be regarded as the extended region of patch. By dividing the feature map into two groups of spatial regions, the dimensionality of the feature map is tripled. The feature information within each group of regions is reduced to half of the original one. To balance the parameter increase due to spatial grouping, we reduce the output dimension to half of the original one. In terms of computation, since the spatial dimension is reduced, the two groups of features can be reasonably matched to reduce the computation. The computational amount of FLOPs before spatial grouping is represented by the following equation 3:

$$FLOPs = C' \times K^2 \times H \times W \times C'' \tag{3}$$

where C' denotes the current number of input channels, C'' denotes the current number of output channels and K is the size of the convolution kernel. After dividing into two groups by space, the computation is represented by the following equation 4:

$$\begin{aligned} FLOPs &= 2C' \times K^2 \times (\alpha \times H) \times (\alpha \times W) \times \frac{C''}{2} \times 2 \\ &= 2\alpha^2 (C' \times K^2 \times H \times W \times C'') \end{aligned} \tag{4}$$

From equations 3 and 4, the amount of calculation is $2\alpha^2$ times the original. To not increase the computational effort, only $2\alpha^2 < 1$ i.e. $\alpha < \frac{\sqrt{2}}{2}$ is required, and hence an upper limit of α less than $\frac{\sqrt{2}}{2}$ is reasonable. Among the values slightly less than $\frac{\sqrt{2}}{2}$, 0.7 is a suitable choice, i.e., $\alpha \in [0.5, 0.7]$. When α=0.5, the calculation is reduced to $\frac{1}{2}$ of the original. When $\alpha = 0.7$, the computation shrinks to the original $\frac{49}{50}$. In summary, the overall computation can be reduced by $\frac{1}{2}$ to $\frac{49}{50}$ times by PSRS.

As shown in Fig. 5 and Fig. 6, the image is uniformly divided into four patches when $\alpha = \frac{1}{2}$, and each patch contains information from the other three patches when $\alpha > \frac{1}{2}$. This method achieves the adaptive patch size adjustment of the network for different image features during the training process to facilitate spatial communication between individual patches while avoiding the destruction of the original continuous features. Such a spatial segmentation method not only facilitates the network to have the ability to extract large granularity information but also can learn continuous features.

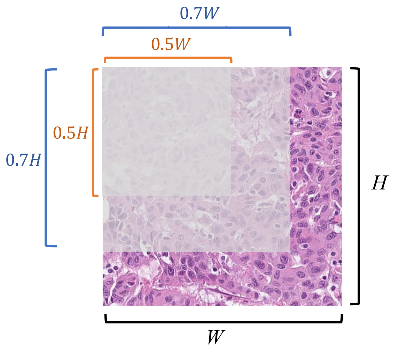
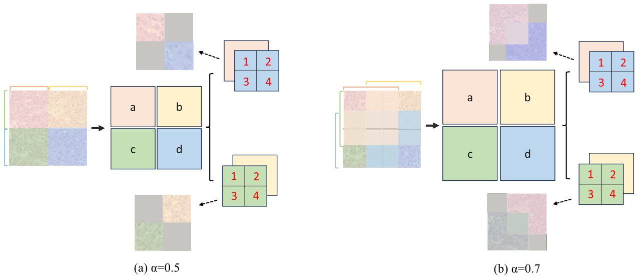


FIGURE 5. The range of values of α.



(a) α=0.5

(b) α=0.7

FIGURE 6. Comparison of effect of α=0.5 and α=0.7.

B. PATCH-WISE FEATURE MATCHING STRATEGY

In the second stage, the network is trained to extract features with smaller local granularity. To achieve this, the large patches are once again sliced and grouped. Each group of feature maps is divided into four smaller patches based on spatial regions, labeled as '1', '2', '3', '4'. Similarly, feature maps '1' and '4' are combined into one group, while '2' and '3' form another group. These smaller patches help the network in extracting smaller granular regions. However, it is important to consider the spatial communication of the small patches at this stage. As the network goes deeper, the size of the patches gradually decreases, resulting in reduced information within it. This lack of effective communication between the groups may lead to a decrease in classification accuracy. Fig. 7 illustrates that when α is approximately 0.5, 'a' and 'd' tend to learn independently, with minimal interaction with the information from 'b' and 'c'.

Based on the above appearances, we designed PFMS to combine the spatial features of the image after reconstructing them. As shown in Fig. 7, the PFMS combines all the '1' and '4' regions to form F_{out1}^{PFMS} after segmenting the features of the large patches again, and at the same time combines all the '2' and '3' regions to form F_{out2}^{PFMS}. This splicing strategy aims to prevent the merging of contiguous regions, allowing small patches to stand alone. By doing so, the initial four patches are reduced to two, enhancing the forward

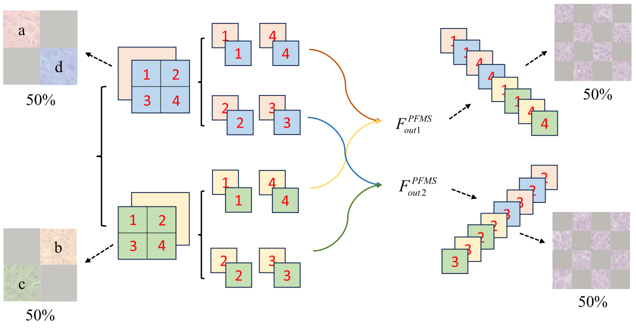


FIGURE 7. **PFMS workflow.**

propagation speed of the network and facilitating information exchange between 'a, b, c, d'. In stage 1, each patch initially contained approximately 50% of the input information. With the addition of PFMS in stage 2, each patch still retains about 50% of the input information. However, the granularity of the patch for each subgroup decreases, making it more extensive and comprehensive in terms of information. This also leads to the gradual merging of local features and their convergence towards global features.

C. PATCH EMBEDDING

The main purpose of patch embedding is to perform linear transformations in order to enhance the expressiveness and perceptual range of the input data. This operation is widely used in the transformer family of models [31], [32], [33]. In our approach, by applying linear transformations to the global features in stage 3, we can further enhance the expressive power of the data. Without feature transformation in stage 1 and stage 2, the original images cannot directly communicate with features. Hence, patch embedding is designed to transform the input data into a linear feature representation by applying different weight matrices W to the input data.

The process of operation still does not confuse the features under local granularity. Taking the output of the feature from PFMS as an example, the patches in F^{PFMS}_{out} are weighted to adjust their importance for adaptive feature fusion and feature selection. The $F^{embedding}_{out}$ in equation 5 is the result after this operation in stage 2.

$$F^{embedding}_{out} = F^{PFMS}_{out} \otimes W^{p}_{out} \tag{5}$$

where p denotes the number of patches in the set of features.

D. PATCH-WISE INFORMATION COMMUNICATION

Progressive learning strategy has been effective in enabling the network to learn features at various levels of granularity. However, these strategies have somewhat weakened the connection between patches. While spatial information transfer still exists, its impact is limited. For instance, when $\alpha = 0.5$ or similar, the feature maps within each group are only influenced by the input of the corresponding patch. As a result of the lack of connection between different patches, the learned features are constrained and may result in information loss. Therefore, we propose enhancing the information interaction between different patches at the channel level.

The before and after comparisons of performing information interaction are shown in Fig. 8. In Fig. 8(a), which is the case without information interaction, it can be seen that there is always no interaction among the three patches after PMEN operates on the input features. In Fig. 8(b) after the information interaction, each patch contains the information of other patches internally but still focuses on its features, and the output features after PMEN are the features after the interaction.

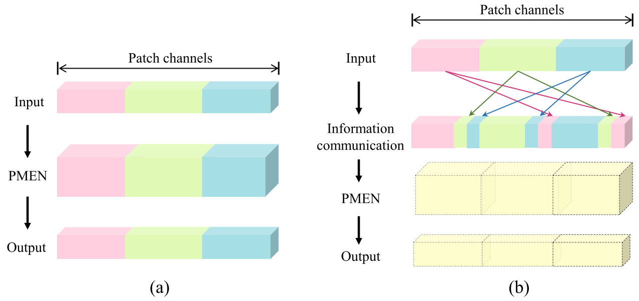


FIGURE 8. **Before and after information interaction.**

Due to the difference in the size and number of patches in stage 1 and stage 2, although the design idea is the same, there are still big differences in the specific design process. We specifically divide the channel communication part for stage 1 and stage 2 into the CISM and GWFM.

1) CHANNEL INTERACTION AND SHIFTING MODULE

In stage 1, PSRS spatially divides the feature map into two groups, each containing two patches. CISM ensures that most of the features within its own patch are included in each group. This is achieved by moving a portion of the channels from the first patch to the back of the last patch, or a portion of the channels from the last patch to the front of the first patch. This operation allows the transfer of information between channels without introducing features from other patches or causing confusion in channel information. The trainable parameter β controls the number of channels moved.

Regarding the value of β, we believe that in the early stages of feature extraction, it is important that the characteristics of each patch dominate and minimize interference from other patches. This helps preserve the uniqueness of each Patch. However, we also recognize the need for information interactions between patches to capture useful information effectively. Therefore, we choose a relatively balanced value of β between 0 and 0.3 to satisfy these requirements in practical applications. In Fig. 9, when $\beta = 0.25$, the number of channels in a patch is shifted by approximately 25%. This means that each group contains 75% of its own features and 25% of features from another patch when convolution is performed again.

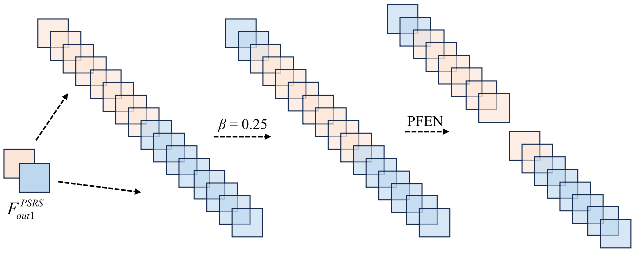


FIGURE 9. Design of CISM.

2) GROUP WEIGHTED FUSION MODULE

In stage 2, the features after PFMS also form two groups, but the number of patches contained in each group becomes eight, so β alone does not achieve mutual communication among the eight patches. We borrowed the idea of ShuffleNet [34], and the two groups of features are Unfolded from 1D to 2D by channel in terms of the number of patches, and then Flattened to 1D after transposition. Fig. 10 shows the feature map F_{out1}^{PFMS} as an example, the number of patches in F_{out1}^{PFMS} is 8, and after unfolding, transposing and flattening, the feature map $shuffle^1$ is formed by mixing all the patches.

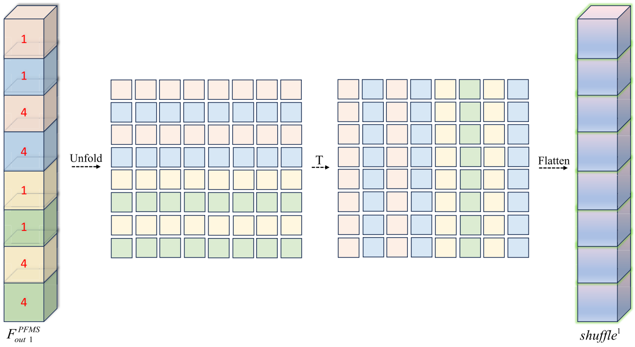


FIGURE 10. F_{out1}^{PFMS} transposed(T) to form a $shuffle^1$.

Unlike ShufflNet, our model requires fully exploring the features of each patch and integrating the information of other patches at the same time, so instead of using $shuffle^1$ directly for learning, we choose to fuse F_{out1}^{PFMS} and $shuffle^1$. Classical feature fusion methods, such as channel splicing, and additive and multiplicative operations, not only increase the parameters and computational complexity of the model but also confuse the information between patches, making it difficult to reflect their relative importance in the model. Therefore, we adopt a grouped weighted summation strategy to fuse F_{out}^{PFMS} and shuffle into a new feature F_{out}^{GWFM} to achieve full interaction between patches. The fusion process is carried out in batch mode, as shown in Fig. 11, where F_{out1}^{PFMS} and $shuffle^1$ are divided into 8 batches according to the Patch boundary, and the cosine similarity of the corresponding batches is subsequently calculated.

In the process of calculating the cosine similarity, we added a trainable parameter λ to increase the error tolerance, and the initial value of λ is 1. This is in order not to change the results of the first-time similarity calculation, and the value of λ is adjusted during the training process to adjust the error of the similarity calculation.

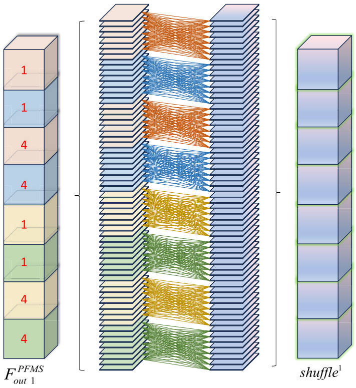


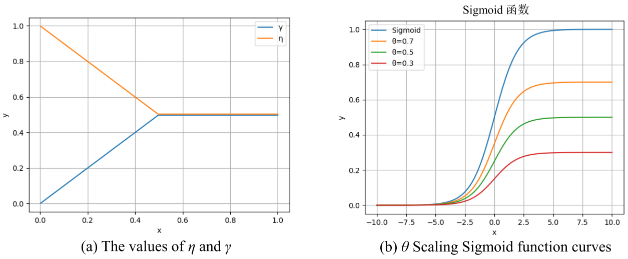
FIGURE 11. F_{out1}^{PFMS} and $shuffle^1$ calculate the cosine similarity.

$$
Similarity = \sum\nolimits_{i=1}^{p} \frac{\sum_{j=1}^{n} (F_{out}^{PFMS\,j} \times shuffle_i^j)}{\sqrt{\sum_{j=1}^{n} (F_{out}^{PFMS\,j})^2} \times \sqrt{\sum_{j=1}^{n} (shuffle_i^j)^2}} \times \lambda \tag{6}
$$

n denotes the dimension of the input features. As in Equation 7, we use two parameters η and γ to control the weight of F_{out}^{PFMS} and $shuffle$ in the output F_{out}^{GWFM} of the GWFM, respectively.

$$
F_{out}^{GWFM} = \sum\nolimits_{i=1}^{p} (\eta_i \times F_{out}^{PFMS} + \gamma_i \times shuffle) \tag{7}
$$

The values of γ are derived by mapping the results of the cosine similarity calculation through a Sigmoid function, $\eta = 1 - \gamma$, and they take the values shown in Fig. 12(a). Since we want to dominate the original features in F_{out}^{PFMS}, we set the upper limit of the value of γ to 0.5, as a way to ensure that γ cannot exceed η.



(a) The values of η and γ

(b) θ Scaling Sigmoid function curves

FIGURE 12. Values of η and γ (left) and Sigmoid function curves under the control of θ (right).

The Sigmoid function is not directly applicable, as the sigmoid curve in Fig. 12(b), when the result of cosine similarity calculation is negative it indicates that the two feature maps are negatively correlated, however, the Sigmoid function still maps the value of *Similarity* to be large in some cases. To solve this problem, we added a trainable parameter θ. As shown in Fig. 12(b), θ is a value between 0.3 and 1.0, $\gamma = f(sigmoid(Similarity), \ \theta)$, which is used to scale the values of γ by compressing the Sigmoid function.

$$\gamma = \begin{cases} \dfrac{\theta}{1+e^{-Similarity}}, & 0 < \dfrac{\theta}{1+e^{-Similarity}} \\ & \leq 0.5; \ (0.5 \leq \theta \leq 1) \\ 0.5, & \dfrac{\theta}{1+e^{-Similarity}} > 0.5 \end{cases} \tag{8}$$

E. STAGE-ADAPTIVE FEATURE ENHANCEMENT MODULE

At different stages of the network, after the first stage of PSRS and CISM and the second stage of PFMS and GWFM, the Patch contains more complex information. The third stage deals with features at the granular level of the original image, which also introduces complexity to the information. In order to quickly capture important information at each stage, we introduce a layer of attentional mechanisms in each stage to enhance the features within. However, traditional attention mechanisms [35], [36], [37] did not meet our specific needs. We not only require enhancing the feature information at different stages but also personalizing the enhancement process based on the characteristics of each stage's features. Therefore, to solve this problem, we propose the SAFEM.

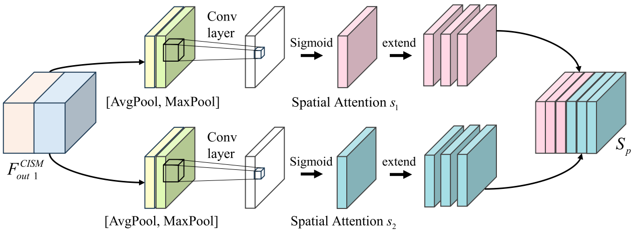


FIGURE 13. SAFEM workflow.

The workflow of SAFEM in Fig. 13 takes the feature $F_{out\,1}^{CISM}$ output from CISM in stage 1 as an example. Firstly, we obtain the channel position corresponding to each patch, and then take the maximum value and average value on the channel of each feature point; after that, we fuse these two results, adjust the number of channels by using the convolution with the number of channels as 1, and then perform one operation with the Sigmoid function. At this point, we obtain the weights s_1 and s_2 of each feature point in the two patches. To reduce the computational complexity, we no longer add the convolution layer to upgrade its dimension, but the s_1 and s_2 are stacked until the number of channels is the same as that of the corresponding patch. Finally, they are all stitched together to form weights of the same dimension as the input features, S_p. As in equation 9, multiplying S_p by the original input feature layer completes the information enhancement for each Patch individually.

$$\begin{aligned} F_{out}^{SAFEM} &= \{F_{out}^{CISM}, F_{out}^{GWFM}\} \otimes \{s_1, s_2, \ldots, s_{2p}\} \\ &= \{F_{out}^{CISM}, F_{out}^{GWFM}\} \otimes S_{2p} \end{aligned} \tag{9}$$

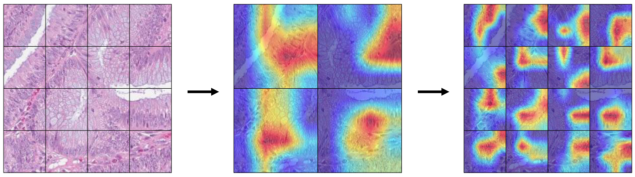


FIGURE 14. Effect of SAFEM on different granularity features.

SAFEM enhances the features of each patch based on its internal characteristics. Fig. 14 illustrates the feature regions enhanced by two local granularity patches after applying SAFEM. Fig. 14(a) represents the input original image, Fig. 14(b) represents the region enhanced based on local large-grained features, and Fig. 14(c) represents the region enhanced based on local small-grained features. It can be observed that the patches in Fig. 14(b) focus on larger epithelial and mesenchymal features, while the patches in Fig. 14(c) focus on densely distributed and less extensive cellular regions. This method of feature enhancement enables the network to accurately capture features at different granularities in colorectal cancer tissues.

F. PROGRESSIVE MULTI-FEATURE EXTRACTION NETWORK

We develop a network called PMEN to address the need for maintaining independence between individual patches in order to perform fine-grained learning for each Patch. This network is specifically designed for the progressive learning method and enables feature extraction.

We choose to utilize the 3×3 convolution as the primary method for feature extraction in our study. This decision is made due to the fact that 3×3 convolutions have fewer parameters compared to larger kernel sizes. By reducing the number of parameters and computational complexity, we are able to mitigate the issue of overfitting. Additionally, stacking and combining multiple 3×3 convolutions allows us to effectively extract multi-granularity features. This method has been successfully employed in various well-known convolutional neural networks such as VGG [38], GoogleNet [39], and ResNet [40], which have demonstrated exceptional performance. Considering that the progressive learning strategy requires learning features at different local granularities and global features consecutively, it often demands significant computational resources. Therefore, we have also taken into account the optimization of parameter amount and computational workload during the network design process. The number of parameters and computation amount for one layer of 3×3 convolution are as follows:

$$Params = C' \times 3^2 \times C'' \tag{10}$$

$$FLOPs = C' \times 3^2 \times h \times w \times C'' \tag{11}$$

h and w are the height and width of the current input, respectively. When a 1×1 convolution is added before and after the 3×3 convolution for lifting and lowering dimensions respectively, the number of parameters and computation are as follows:

$$\begin{aligned}Params &= C' \times 1^2 \times \frac{C''}{2} + \frac{C''}{2} \times 3^2 \times \frac{C''}{2}\\ &\quad + \frac{C''}{2} \times 1^2 \times C'' = \frac{2C' + 11C''}{4} \times C'' \qquad (12)\\ FLOPs &= C' \times 1^2 \times h \times w \times \frac{C''}{2}\\ &\quad + \frac{C''}{2} \times 3^2 \times h \times w \times \frac{C''}{2}\\ &\quad + \frac{C''}{2} \times 1^2 \times \frac{h}{2} \times \frac{w}{2} \times C''\\ &= \frac{4C' + 19C''}{8} \times C'' \times h \times w \qquad (13)\end{aligned}$$

By looking at equations 10 and 12, we find that the number of parameters after adding the 1×1 convolution is $\frac{2C'+11C''}{36C'}$ times the original number. To reduce the number of parameters, it is necessary to satisfy $\frac{2C'+11C''}{36C'} < 1$. The number of input channels and the number of output channels conform to the relation $\frac{C''}{C'} < \frac{34}{11}$, i.e., when the output channels are approximately three times the number of input channels or less. Observing equations 11 and 13, we find that the amount of computation is $\frac{4C'+19C''}{72C'}$ times the original amount, and the same calculation shows that the amount of computation is reduced when the number of input channels and the number of output channels conform to the relationship $\frac{C''}{C'} < \frac{68}{19}$, i.e., the number of output channels is approximately 3.6 times or less than the number of input channels.

Based on the above analysis and the patch features output from each stage, we design a method called patch-wise convolution that performs convolution according to the features of each patch separately. Specifically, the input features are divided into p groups for convolution separately based on the number of patches p, which can both reduce computational resources and target each block of patch features for targeted learning. Fig. 15 shows an example diagram of patch-wise convolution, assuming that the number of input patches is 2 and the number of its input channels is C. Fig. 15(a) is the first 1 × 1 patch-wise convolution for halving the dimensionality to reduce the number of parameters and computation generated by the subsequent computation. Fig. 15(b) is the 3 × 3 patch-wise convolution, which is the main part of the feature extraction, and its input and output channels are kept the same. Fig. 15(c) is the second 1×1 patch-wise convolution, which is used to reduce the number of channels to the dimension that is originally intended to be output. It can be seen that in the design for the channel aspect, the output channels of all the convolutional layers are kept at twice or less the number of input channels, thus conforming to the reduced number of parameters and computations.

The number of parameters and the amount of computation for patch-wise convolution are represented by equations 14 and 15:

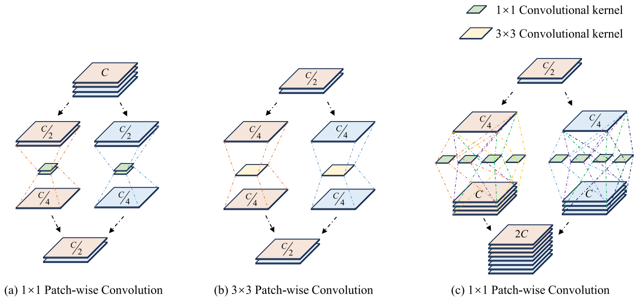

1×1 Convolutional kernel
3×3 Convolutional kernel
C
C/2
C/4
2C
(a) 1×1 Patch-wise Convolution
(b) 3×3 Patch-wise Convolution
(c) 1×1 Patch-wise Convolution

FIGURE 15. **Patch-wise Convolution.**

$$\begin{aligned}Params &= (\frac{2\frac{C'}{p} + 11\frac{C''}{p}}{4} \times \frac{C''}{p}) \times p\\ &= (\frac{2C' + 11C''}{4} \times C'') \times \frac{1}{p} \qquad (14)\\ FLOPs &= (\frac{4\frac{C'}{p} + 19\frac{C''}{p}}{8} \times \frac{C''}{p} \times h \times w) \times p\\ &= (\frac{4C' + 19C''}{8} \times C'' \times h \times w) \times \frac{1}{p} \qquad (15)\end{aligned}$$

Patch-wise convolution reduces the number of parameters and computation to $\frac{1}{p}$ of the previous one.

To enhance the performance of the model, We divide PMEN Block into Part1 and Part2 and incorporate deformable convolution [41] in Part1. This convolution allows the network to downsample the feature map based on its morphological features. Deformable convolution involves adding an offset, learned by the model, to the normal convolution. In order to better accommodate the characteristics of colorectal cancer, we carefully analyzed the images and observed that the internal mesenchyme often exhibited a strip-like structure resembling papillae and villi. Additionally, the cells tended to deform in a similar manner after being extruded. Therefore, we modified the deformable convolution to account for this feature by increasing the offset in the transverse and longitudinal directions. This adjustment enables the overall offset to quickly adapt to the structure of colorectal cancer. This modified convolution, referred to as patch-wise deformable strip convolution (patch-wise DS convolution), is still applied individually to each patch.

$$y(p_0) = \sum\nolimits_{p_n \in R} w(p_n) * x(p_0 + p_n + \Delta\widetilde{p_n}) \qquad (16)$$

p_0 is each point on the output feature map, $y(p_0)$ is the specific value of each point on the output feature map, p_n represents the offset of each point in the convolution kernel concerning the centroid, and $w(p_n)$ represents the weight of the corresponding position on the convolution kernel. $x(p_0 + p_n)$ is the specific value of each point on the output feature map that corresponds to the region of convolution sampling on the input feature map. $\Delta\widetilde{p_n} = \Delta p_n + \tau$, where Δp_n denotes the offset against the input feature map,

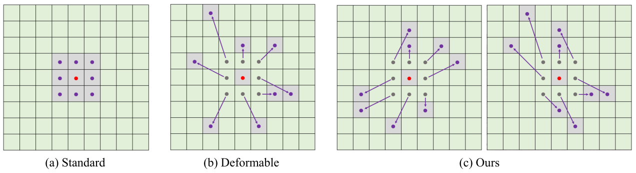


FIGURE 16. **Standard Convolution, Deformable Convolution and Patch-wise DS Convolution.**

and τ denotes the additional offset in the direction of the horizontal and vertical axes. In deformable convolution, the output needs to learn the parameter doubled in the convolution kernel, with the front denoting the horizontal axis coordinates and the back denoting the vertical axis coordinates. τ is a learnable parameter that is used to add additional offsets in the horizontal and vertical axis directions. As can be seen in Fig. 16, compared to the standard convolution Fig. 16(a) and deformable convolution Fig. 16(b), the deformable strip convolution in Fig. 16(c) exhibits two different convolutional trends based on the morphological characteristics of the CRCs pointed out above, which can be adaptively selected during the learning process.

To gain a better understanding of the impact of patch-wise DS convolution, we selected two representative samples that align well with the two forms of patch-wise DS convolution mentioned earlier. As depicted in Fig. 17, we applied standard convolution, deformable convolution, and patch-wise DS convolution to the two CRC images Fig. 17(a). We observed that when deformable convolution Fig. 17(c) is used, the network is able to more effectively incorporate cellular features compared to standard convolution Fig. 17(b). Conversely, our patch-wise DS convolution Fig. 17(d) facilitated the network in capturing the structural characteristics of the CRC in an exceptional manner, resulting in remarkably clear mesenchyme and cellular morphology.

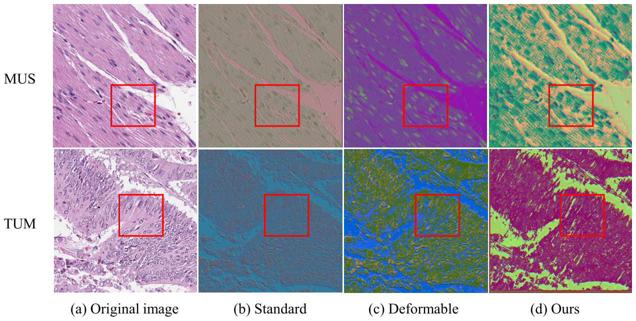


FIGURE 17. **Comparison of the effects of Standard convolution, Deformable convolution and Patch-wise DS convolution.**

In summary, 1 × 1 patch-wise convolution, 3 × 3 patch-wise convolution, and 3 × 3 patch-wise DS convolution together form the basic feature extraction module. As shown in Fig. 18, the PMEN Block consists of Part1 and Part2, which are used for downsampling and feature extraction tasks, respectively. The whole PMEN uses only the 3 × 3 patch-wise DS convolution with step size 2 in Part1 for downsampling and also incorporates the normalization and ReLU activation functions.

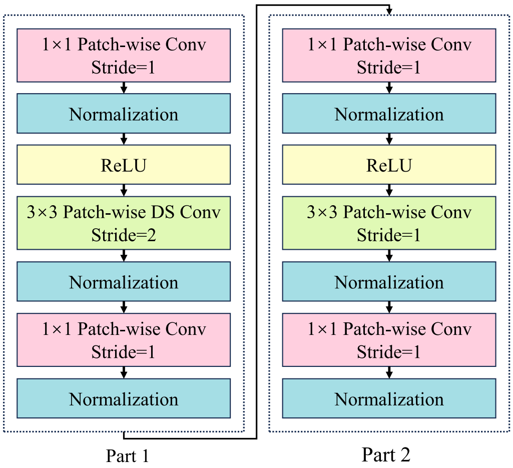


FIGURE 18. **PMEN Block.**

The output images from stage 1 are divided into two groups, each containing two patches. To ensure that the network could learn the features of these patches independently, the parameter p in PMEN is set to 2. In stage 2, the number of patches in each group is increased to 8. The parameter p in PMEN is then set to 8 to allow the network to learn 8 small patches independently. The design of PMEN ensures that the information between the patches is not confused, while also reducing the number of references and computational effort. After learning about the two local granularities, PMEN is still used in stage 3 to learn the features at the granularity level of the original map and combine the information from the two local granularities in the original map. Since stage 3 is performed in the original map, it focuses on the global features of CRC under the guidance of PMEN.

IV. EXPERIMENT

A. DATASETS

To further validate the effectiveness of the proposed method, we selected three public datasets for our experiments. These are 11977 image patches of hematoxylin and eosin-stained human colorectal cancer histological samples [14], 100 colorectal adenocarcinoma images [42], and 5000 colorectal cancer histological images obtained by Aperio digital pathology scanner [43].

The 11977 image patches of the first dataset contain 3 major categories that can be subdivided into 6 subcategories. These are 3977 sparse non-tumor tissue (ADIMUC); 4000 dense non-tumor tissue (STRMUS); and 4000 tumor tissue (TUMSTU). All images are PNG images of 512 × 512 size. The type of the second dataset contains a colorectal adenocarcinoma and the images have a larger size of 1000 × 1000 pixels. We use the first dataset as a training set, and due to the large number and relatively small number of types in this dataset, to further increase the difficulty of classification

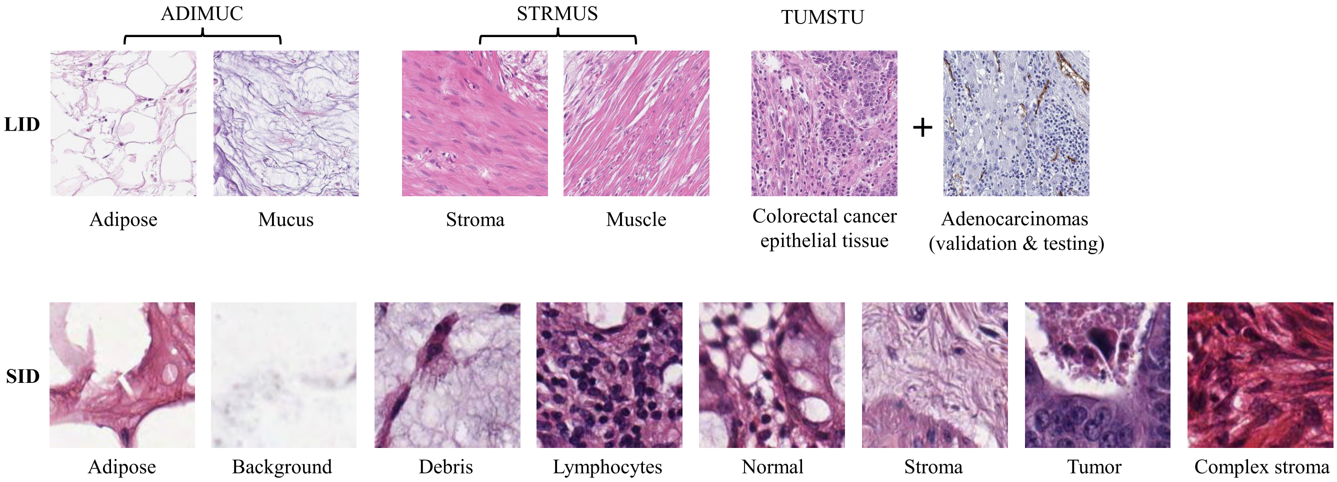


FIGURE 19. Datasets.

as well as to test the generalization ability of the model, the second dataset is cropped to 400 images of 512 × 512 size and merged into the TUMSTU of the validation set and the test set, respectively. the two larger-sized datasets are referred to as Large Image Dataset (LID), such as the LID in Fig. 19.

The third dataset contains 5000 histological images of colorectal cancer with a size of 150 × 150 pixels. This dataset is acquired with the Aperio Digital Pathology Scanner and has more categories and smaller image sizes than the first two datasets, thus having a higher classification difficulty. By analyzing this dataset, the aim is to assess whether the proposed method can show good performance in the face of greater classification difficulty. The dataset has 625 images of each type, including eight colorectal cancer types, namely adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM) and complex stroma (COMP). Call this dataset Small Image Dataset (SID) as in Fig. 19.

B. EXPERIMENTAL SETTINGS AND EVALUATION METRICS

We choose the efficient deep learning framework PyTorch to implement our proposed classification method, and the training and testing of the models are performed in a GPU environment with an NVIDIA GeForce RTX 3090 graphics card. The LID is randomly cropped to 224×224 size in training and the SID is not cropped due to its smaller size, and data enhancement operations such as random flipping, color enhancement, greyscaling, and normalization are applied to both types of data. The classification loss is calculated using the cross-entropy loss function together with label smoothing, stochastic gradient descent is used as the optimizer, the initial learning rate is set to 0.001, the weights are decayed to 5e-4 and the momentum is 0.9.

In this paper, all experiments are trained with 200 epochs, and the classification performance of the model is evaluated using several commonly used metrics in medical image classification, namely Accuracy, Precision, Recall, and F1-score. Accuracy represents the proportion of correctly classified samples of the model to the total number of samples. Precision denotes the proportion of all positives predicted by the model that is correctly predicted; Recall denotes the proportion of all true positives that are correctly predicted by the model; F1-score combines the results of Precision and Recall outputs. The meaning of each indicator is as in equations 17, 18, 19 and 20:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{20}$$

where TP indicates true positive, FP indicates false positive, TN indicates true negative, and FN indicates false negative.

C. PARAMETER DETAILS

The progressive learning strategy is an integral part of PMFF. In Tables 1, 2, and 3, we provide the general structure of PMFF for learning three kinds of granular information in three stages. This includes detailed data on the variation of output size and number of channels, the number of spatial groupings, and the number of patches in each group. The table assumes that the size of the input image is 224 × 224 and its downsampling operation occurs only in Part1 of PSRS, PFMS, and PMEN Block. Patch embedding, CISM, and SAFEM do not change the dimension of the feature maps. PMEN consists of three PMEN Blocks stacked on top of each other.

TABLE 1. Overall composition of Stage 1.

Stage 1	Output size	Output channels	group	patch/group
PSRS	112×112	6	2	2
Patch Embedding	112×112	6	2	2
CISM	112×112	6	2	2
SAFEM	112×112	6	2	2
PMEN Block	56×56	8	2	2
PMEN Block	28×28	16	2	2
PMEN Block	14×14	32	2	2

TABLE 2. Overall composition of Stage 2.

Stage 2	Output size	Output channels	group	patch/group
PSRS	112×112	6	2	2
PFMS	56×56	12	2	8
Patch Embedding	56×56	32	2	8
GWFM	28×28	32	2	8
SAFEM	28×28	32	2	8
PMEN Block	14×14	64	2	8
PMEN Block	7×7	128	2	8
PMEN Block	4×4	256	2	8

TABLE 3. Overall composition of Stage 3.

Stage 3	Output size	Output channels	group	patch/group
Patch Embedding	224×224	3	1	1
SAFEM	224×224	3	1	1
PMEN Block	112×112	64	1	1
PMEN Block	56×56	128	1	1
PMEN Block	28×28	256	1	1

TABLE 4. Comparison of number of parameters and computation before and after considering computational complexity.

Optimization item	Params	FLOPs
Before optimization	6.21 M	1.12 G
PSRS & optimization	6.21 M	1.11 G
PMEN & optimization	4.88 M	0.42 G
PMFF (ours)	4.87 M	0.41 G

D. COMPUTATIONAL COMPLEXITY EXPERIMENTS

The progressive learning strategy performs multiple learning of features, which improves performance but also involves computational complexity. The computational complexity optimization of the algorithm is considered during the segmentation operation of PSRS and the design of PMEN. In order to visually present the effect comparison before and after the computational complexity optimization, Table 4 takes a 224×224 pixel RGB image as an example to exhaustively show the change of the number of parameters and the computational amount. Where PSRS is the case assuming α=0.7, after considering the computational complexity at the design time, although the number of parameters is not reduced, the computational amount is reduced by 0.01G. On the other hand, after considering the computational complexity of PMEN, both the number of parameters and the computational amount are significantly reduced. The combination of the two forms the final PMFF, where both the number of parameters and the computational volume are substantially optimized.

TABLE 5. Comparison with progressive learning strategy used for classification tasks.

Method	Params	FLOPs
PMG (Resnet-18) [22]	30.87 M	7.09 G
PMG (Resnet-34) [22]	40.98 M	8.94 G
PMG (Resnet-50) [22]	44.74 M	9.36 G
PJGC-Net [24]	29.45 M	8.75 G
PMFF (ours)	4.87 M	0.41 G

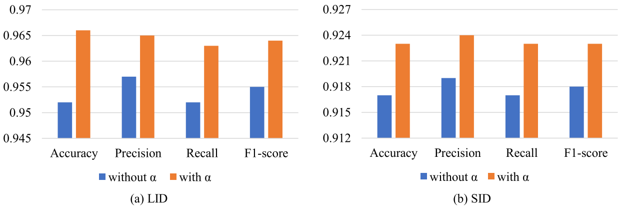
We compare this with some progressive learning strategies PMG [22] and PJGC-Net [24] for classification tasks in terms of computational complexity. Where PMG provides several base models, the higher the number of layers in the network, the higher the complexity will be, the authors used resnet-50 as the default base model. As can be seen from Table 5, the progressive learning strategy we designed is much lower than the existing progressive learning strategies both in terms of the parameters and FLOPs.

E. ABLATION STUDY

1) PARAMETER SETTINGS

In the method section, we detail some of the important parameters applied in the proposed method and emphasize their role in feature extraction. Next, we will further verify whether these elements are indeed effective in improving the performance of the model through experimental results. The experiments use two types of datasets as well as four evaluation metrics.

The performance comparison results after the introduction of α in the experiments are presented in Fig. 20. Fig. 20(a) and Fig. 20(b) show the experimental results on two types of datasets, respectively. It is evident from the figure that the introduction of α leads to performance improvement in all four evaluation metrics. This result intuitively confirms the effectiveness of learning continuous features for enhancing performance.



(a) LID

(b) SID

FIGURE 20. Before and after the use of α on two types of datasets.

Moving on, we focus on the experimental validation of β and cosine similarity in Fig. 21. Firstly, we test the role of β under CISM, and the results demonstrate that the introduction of β further enhances the system performance. Similarly, the use of cosine similarity under GWFM for feature fusion shows significant performance enhancement. This suggests that by cleverly incorporating elements such as β and cosine similarity in the process of message communication, the model exhibits superior performance in both aspects.

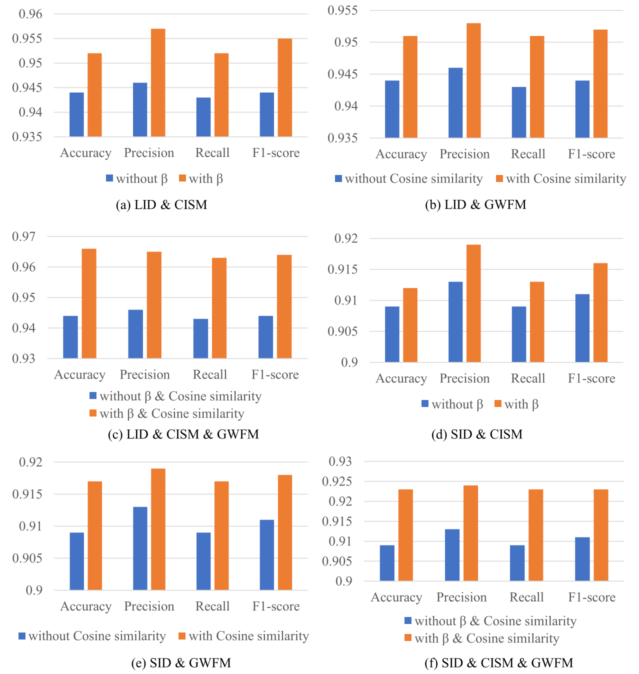


FIGURE 21. *β* **and cosine similarity before and after use.**

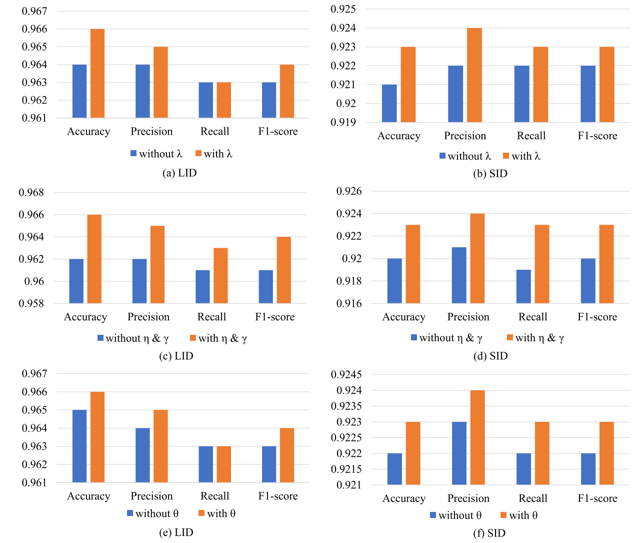


FIGURE 22. λ, η, γ, **and** θ **before and after use.**

When using cosine similarity for feature fusion, the parameters λ, η, γ, and θ are added. As shown in Fig. 22, λ serves to increase the error tolerance of the cosine similarity computation, while θ is used to scale the Sigmoid function curves, which play a role even though their enhancement is relatively small. η and γ, which exist as a whole, serve to ensure that patch's original features make up a major portion of the features, which are validated from the evaluation metrics and are verified to be effective.

In the design of PMEN, a parameter τ is added to improve the deformable convolution. τ serves to adapt the deformable convolution to the morphological features of CRC faster. From Fig. 23, it can be seen that the addition of τ achieves some effect, and all four indexes are slightly improved.

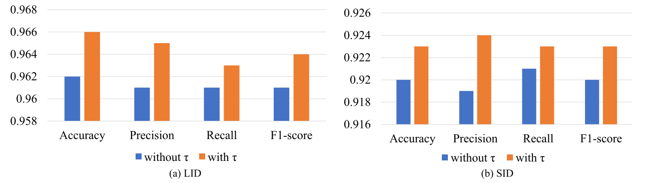


FIGURE 23. τ **before and after joining.**

2) ABLATION EXPERIMENT OF PROGRESSIVE LEARNING STRATEGY

In this subsection of the experiment, we compare the learning effects of the three stages of progressive learning, namely CISM, GWFM, SAFEM, and PMEN. Stage 3 is an indispensable part as it combines the information of two local granularities to better preserve the spatial structure features of the CRC on the original map. The experimental results in Table 6 show that when stage 3 is used alone to learn the global features of the original image, the overall performance is approximately 92.8% and 85.9%, respectively. However, when stage 1 is added to train the ability of the network to capture information at a higher level of granularity., there is an improvement of around 1.3% and 2% in performance. Similarly, when stage 2 is added, the network gains the ability to extract smaller granularity information, resulting in higher performance compared to using only stage 3 for global feature extraction. The best performance is achieved when all three stages (stage 1, stage 2, and stage 3) are used together, allowing the network to effectively fuse global features and features at different local granularities. This leads to an improvement of about 3.7 and 6.4 percentage points over using only stage 3.

3) VISUAL ANALYSES AT ALL STAGES

To assess the impact of progressive learning, we utilize a heatmap [44] to visualize the CRC tissues in Fig. 24. The results show that during the training process of stage 1, the network focuses on features with larger granularity, such as the cup cells in the stomach tumor along with the surrounding glandular cells and the tumor tissue in the colorectal tumor. In contrast, the focus of the network shifts to features of smaller granularity during stage 2 training, such as the mesenchyme in adipose and the scattered nuclei in stomach tumors. Stage 3 further emphasizes global features, as it simultaneously learns two parts with different granularity. For example, in the stomach tumor, it learns both the cup-shaped cells and the scattered cells. The final concatenation represents the most discriminative region of interest after combining the features learned in the three stages. This region typically exhibits continuous and non-redundant features.

TABLE 6. Performance comparison of progressive learning stages.

Method			Accuracy(%)		Precision(%)		Recall(%)		F1-score(%)	
Stage1	Stage2	Stage3	LID	SID	LID	SID	LID	SID	LID	SID
		✓	92.8	84.7	92.5	84.6	93.0	84.7	92.8	84.6
✓		✓	94.1	86.5	93.9	87.0	94.4	86.5	94.1	86.7
	✓	✓	93.9	86.3	93.6	87.0	94.2	86.3	93.9	86.7
✓	✓	✓	96.6	92.3	96.5	92.4	96.3	92.3	96.4	92.3

TABLE 7. Stage 1 ablation experiment results.

Method	Accuracy(%)		Precision(%)		Recall(%)		F1-score(%)	
	LID	SID	LID	SID	LID	SID	LID	SID
Backbone1	94.2	86.6	94.1	86.8	93.9	86.6	94.0	86.7
Backbone1+CISM	94.6	87.5	94.5	88.1	94.5	87.5	94.5	87.9
Backbone1+SAFEM	95.6	87.5	95.6	88.4	95.3	87.5	95.4	87.8
PMFF (ours)	96.6	92.3	96.5	92.4	96.3	92.3	96.4	92.3

TABLE 8. Stage 2 ablation experiment results.

Method	Accuracy(%)		Precision(%)		Recall(%)		F1-score(%)	
	LID	SID	LID	SID	LID	SID	LID	SID
Backbone2	94.4	87.3	94.3	87.4	94.3	87.3	94.3	87.3
Backbone2+GWFM	95.0	88.2	95.0	89.2	94.8	88.2	94.9	88.7
Backbone2+SAFEM	95.3	89.2	95.1	90.8	95.1	89.2	95.1	90.0
PMFF (ours)	96.6	92.3	96.5	92.4	96.3	92.3	96.4	92.3

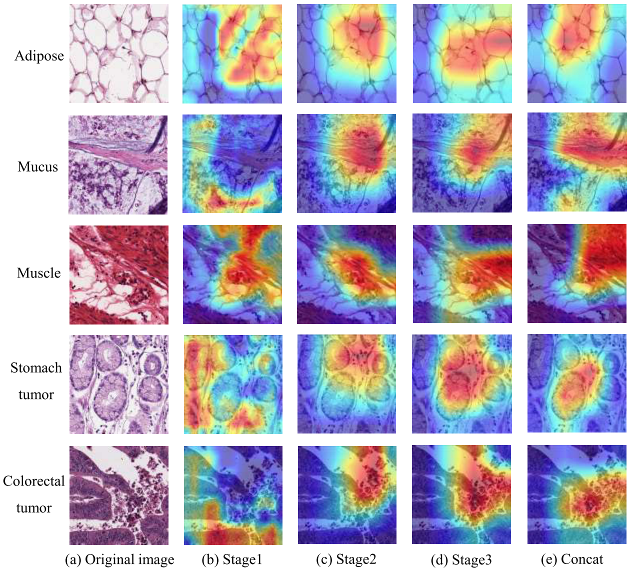


FIGURE 24. Comparison of results by stage.

4) ABLATION EXPERIMENTS AT VARIOUS STAGES

Given the previous experiments that have demonstrated the superiority of progressive learning, we proceeded to conduct ablation experiments on each stage within the progressive framework. In an ablation experiment, a specific stage is isolated while the modules from the remaining stages are retained. We first conduct ablation experiments on stage 1, where PSRS, patch embedding, and the feature extraction network PMEN are employed as Backbone1. The experimental results, presented in Table 7, reveal that Backbone1 achieves an accuracy of 94.2% and 86.6% on the two types of datasets, respectively. When CISM is introduced, there is increased interaction of information between each patch, leading to improved performance. Furthermore, the overall performance is further enhanced by more than 1% with the addition of SAFEM, particularly in capturing larger local granularity features like tumor tissue.

In the second stage of the experiment, Backbone2, consisting of PSRS, PFMS, patch embedding, and PMEN, is utilized. The experimental results are presented in Table 8, demonstrating that Backbone2 achieves an accuracy of 94.4% and 87.3% on the two types of datasets, respectively. Furthermore, when GWFM is incorporated, the local smaller granularity features, which initially lack sufficient information, are effectively interacted with. This interaction leads to an overall performance improvement of 0.6% and 1.25%. Additionally, the inclusion of SAFEM to enhance smaller-grained local features, such as piles, further enhances the overall performance by 0.8% and 2.5%.

In stage 3, PMEN serves as the Backbone3 to evaluate the impact of patch embedding and SAFEM. The experimental results are presented in Table 9, indicating that Backbone3 achieves an accuracy of 94.5% and 89.4% on the two types of datasets, respectively. The inclusion of patch embedding leads to a slight improvement in performance, validating the continued usefulness of linear transformation of global features. Furthermore, the incorporation of SAFEM enhances the global features of the image, resulting in an improvement of over one percentage point in performance.

TABLE 9. Stage 3 ablation experiment results.

Method	Accuracy(%)		Precision(%)		Recall(%)		F1-score(%)	
	LID	SID	LID	SID	LID	SID	LID	SID
Backbone3	94.5	89.4	94.2	90.0	94.2	89.4	94.2	89.7
Backbone3+Parch Embedding	94.7	89.7	94.4	90.5	94.5	89.7	94.5	90.1
Backbone3+SAFEM	95.4	91.5	95.5	91.9	95.2	91.5	95.3	91.7
PMFF (ours)	96.6	92.3	96.5	92.4	96.3	92.3	96.4	92.3

TABLE 10. Comparison of results with some existing CRC classification methods on two types of datasets.

Method	Accuracy(%)		Precision(%)		Recall(%)		F1-score(%)	
	LID	SID	LID	SID	LID	SID	LID	SID
SN Macenko [45]	74.9	64.2	77.5	66.5	74.9	64.2	76.2	65.3
SN Vahadane [46]	76.6	69.4	77.4	70.8	76.6	69.4	77.0	70.1
LoFGAN [15]	77.0	69.7	79.2	70.9	77.0	69.7	78.1	70.3
XM-GAN [16]	81.3	72.5	82.7	73.8	81.3	72.5	82.0	73.1
CNN-5B [12]	83.1	77.3	83.2	78.4	83.1	77.3	83.2	77.8
FE-SVM(RBF) [14]	89.5	87.4	91.0	88.8	89.5	87.4	90.3	88.1
IMPaSh [13]	93.1	89.2	93.5	90.8	93.2	89.2	93.3	90.0
DenseNet169-SVM [11]	95.7	92.1	96.1	91.8	95.6	92.1	95.8	91.9
PJGC-Net [24]	95.9	92.1	96.0	92.0	95.9	92.1	95.9	92.0
Federated learning [17]	95.9	92.0	96.3	92.1	96.0	92.0	96.1	92.0
PMG [22]	96.2	92.1	96.1	92.0	96.1	92.1	96.1	92.1
DMDA [18]	96.3	92.2	96.3	92.4	96.1	92.1	96.2	92.2
PMFF (ours)	96.6	92.3	96.5	92.4	96.3	92.3	96.4	92.3

F. COMPARISONS WITH EXISTING CLASSIFICATION METHOD

Table 10 compares the four evaluated metrics with several existing methods for classifying colorectal cancer [11], [12], [13], [14], [15], [16], [17], [18], [45], [46] and progressive learning strategy for classification tasks [22], [24]. PMFF achieved the accuracy of 96.6% and 92.3% and the f1-score of 96.4% and 92.3% on the two types of datasets. It outperformed the second-place finisher by 0.3% in accuracy and 0.1% in f1-score. Additionally, it outperformed the existing methods listed in the table overall, with improvements of 0.2% and 0.1% in f1-score, and 0.2% in precision on LID. In SID, the precision of DMDA is equal to that of the proposed method, indicating the superiority of focusing on local and global distribution alignment used by DMDA. Slightly inferior to PMG in recall, illustrating the superiority of progressive learning strategy and multi-granularity feature fusion.The proposed method not only accurately classifies images but also maintains a good balance between accuracy and recall. PMFF demonstrates consistent performance across all evaluation metrics, indicating that the fusion of global features and different local granularity features ensures stable and reliable classification.

The data in experiments will inevitably be affected by degradation, noise, or variability, leading to deviations in classification accuracy for existing and proposed methods. To mitigate these factors, we utilized various public datasets for our experiments, yet encountered challenges. Hong et al. [47] simulated spectral variations in different environments and suggested solutions to address them. In future research, we plan to leverage the methods outlined in this paper to explore ways to minimize the impact of data degradation, noise, variability, etc., to enhance the robustness and generalization capabilities of classification models.

V. CONCLUSION

In this paper, we propose a novel method called PMFF for the classification task of colorectal histopathological images. Our method adopts a progressive learning strategy, which enables the model to learn different granularity features and fuse information from different granularity levels in the final stage. PMEN is specifically designed to capture the morphological features of CRCs in different stages. It effectively addresses the challenge of capturing complex morphological features. Additionally, CISM and GWFM facilitate information transfer between patches to compensate for the information transfer insufficiency caused by PMEN, thereby enhancing the model's learning ability. SAFEM is employed to expedite the model's ability to locate important information, thereby improving the learning efficiency. Experimental results on multiple datasets demonstrate the effectiveness of PMFF. Furthermore, our method outperforms existing CRC classification methods in terms of several evaluation metrics, providing a reliable solution for accurate CRC classification.

Accurate colorectal cancer (CRC) classification is crucial in clinical practice for optimizing diagnosis and treatment pathways and achieving precision medicine. It aids in providing more precise diagnoses, preventing treatment delays or over-treatment resulting from misclassification. Furthermore, accurate CRC classification can alleviate the workload of healthcare professionals and enhance work efficiency. The CRC classification method introduced in this study is innovative in its approach and design, offering a fresh theoretical perspective and practical approach to pathology image classification.

Although our method shows better performance in classification tasks targeting CRC, there are still some limitations. First, the model relies on labeled data for training data. Second, the computational resources consumed by stage 3 are not as light as those of stage 1 and stage 2. In the future, we will focus on improving the model from these two aspects. On the one hand, we will implement a semi-supervised approach to alleviate the problem of the lack of labeled pathology images. On the other hand, we will reduce the computational overhead to improve the classification efficiency of pathology images for better application in medical diagnosis.

REFERENCES

[1] Y. Xi and P. Xu, "Global colorectal cancer burden in 2020 and projections to 2040," *Translational Oncol.*, vol. 14, no. 10, Oct. 2021, Art. no. 101174.

[2] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X.X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.

[3] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 3, 2024, doi: 10.1109/TPAMI.2024.3362475.

[4] D. Hong, C. Li, B. Zhang, N. Yokoya, J. A. Benediktsson, and J. Chanussot, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in Earth observation," *Innov. Geosci.*, vol. 2, no. 1, 2024, Art. no. 100055.

[5] M. Jahanifar, A. Shepard, N. Zamanitajeddin, R. M. S. Bashir, M. Bilal, S. A. Khurram, F. Minhas, and N. Rajpoot, "Stain-robust mitotic figure detection for the mitosis domain generalization challenge," in *Proc. MICCAI*, Mar. 2022, pp. 48–52.

[6] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2845–2856, Oct. 2021.

[7] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, and J. van der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 160–163.

[8] X. Wu, Y. Li, X. Chen, Y. Huang, L. He, K. Zhao, X. Huang, W. Zhang, Y. Huang, Y. Li, M. Dong, J. Huang, T. Xia, C. Liang, and Z. Liu, "Deep learning features improve the performance of a radiomics signature for predicting Kras status in patients with colorectal cancer," *Academic Radiol.*, vol. 27, no. 11, pp. e254–e262, Nov. 2020.

[9] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Phys. Technol.*, vol. 10, no. 3, pp. 257–273, Jul. 2017.

[10] S. Poudel, Y. J. Kim, D. M. Vo, and S.-W. Lee, "Colorectal disease classification using efficiently scaled dilation in convolutional neural network," *IEEE Access*, vol. 8, pp. 99227–99238, 2020.

[11] E. F. Ohata, J. V.S. D. Chagas, G. M. Bezerra, M. M. Hassan, V.H. C. de Albuquerque, and P. P. R. Filho, "A novel transfer learning approach for the classification of histological images of colorectal cancer," *J. Supercomput.*, vol. 77, no. 9, pp. 9494–9519, Feb. 2021.

[12] V. Rachapudi and G. Lavanya Devi, "Improved convolutional neural network based histopathological image classification," *Evol. Intell.*, vol. 14, no. 3, pp. 1337–1343, Sep. 2021.

[13] T. T. L. Vuong, Q. D. Vu, M. Jahanifar, S. Graham, J. T. Kwak, and N. Rajpoot, "IMPaSh: A novel domain-shift resistant representation for colorectal cancer tissue classification," in *Proc. ECCV*, Feb. 2023, pp. 543–555.

[14] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I.Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLOS Med.*, vol. 16, no. 1, Jan. 2019, Art. no. e1002730.

[15] Z. Gu, W. Li, J. Huo, L. Wang, and Y. Gao, "LoFGAN: Fusing local representations for few-shot image generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8443–8451.

[16] A. Kumar, A. K. Bhunia, S. Narayan, H. Cholakkal, R. M. Anwer, J. Laaksonen, and F. S. Khan, "Cross-modulated few-shot image generation for colorectal tissue classification," in *Proc. ICCV*, Oct. 2023, pp. 128–137.

[17] M. Nergiz, "Federated learning-based colorectal cancer classification by convolutional neural networks and general visual representation learning," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 3, pp. 951–964, May 2023.

[18] W. Yu, N. Xu, N. Huang, and H. Chen, "Bridging the gap: Geometry-centric discriminative manifold distribution alignment for enhanced classification in colorectal cancer imaging," *Comput. Biol. Med.*, vol. 170, Mar. 2024, Art. no. 107998.

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[20] N. Ahn, B. Kang, and K.-A. Sohn, "Image super-resolution via progressive cascading residual network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 904–9048.

[21] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[22] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y. Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Proc. 16th Eur. Conf. Comput. Vis.*, vol. 12365, Aug. 2020, pp. 153–168.

[23] T. Zhang, D. Chang, Z. Ma, and J. Guo, "Progressive co-attention network for fine-grained visual classification," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.

[24] Z. Cao and W. Jia, "Fine-grain classification method of non-small cell lung cancer based on progressive jigsaw and graph convolutional network," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vision.*, Dec. 2023, pp. 409–420.

[25] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5152–5161.

[26] C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, and A. L. Yuille, "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1910–1919.

[27] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2224–2233.

[28] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.

[29] B. Ren, Y. Liu, Y. Song, W. Bi, R. Cucchiara, N. Sebe, and W. Wang, "Masked jigsaw puzzle: A versatile position embedding for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20382–20391.

[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[33] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer V2: Scaling up capacity and resolution," 2021, *arXiv:2111.09883*.

[34] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[35] J. Hu, L. Shen, and G. Sun, "SENet squeeze-and-excitation networks," in *Proc. CVPR*, Jun. 2018, pp. 7132–7141.
[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[37] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[42] J. N. Kather, A. Marx, C. C. Reyes-Aldasoro, L.R. Schad, F. G. Zöllner, and C.-A. Weis, "Continuous representation of tumor microvessel density and detection of angiogenic hotspots in histological whole-slide images," *Oncotarget*, vol. 6, no. 22, pp. 19163–19176, Aug. 2015.
[43] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L.R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner, "Multi-class texture analysis in colorectal cancer histology," *Sci. Rep.*, vol. 6, no. 1, p. 27988, Jun. 2016.
[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[45] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE Int. Symp. Biomed. Imaging, From Nano Macro*, Jun. 2009, pp. 1107–1110.
[46] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans. Med. Imag.*, vol. 35, no. 8, pp. 1962–1971, Aug. 2016.
[47] D. Hong, N. Yokoya, J. Chanussot, and X.X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

ZHENGGUANG CAO was born in Henan, China, in 1999. He is currently pursuing the master's degree in computer technology with Ningxia University, Yinchuan, China. His primary research interest includes medical image processing and analysis.

WEI JIA received the Ph.D. degree in computer application technology from Northwest University, Xi'an, China, in 2019. He is currently an Associate Professor with the School of Information Engineering, Ningxia University. His current research interests include medical image processing and computational pathology.

HAIFENG JIANG received the Ph.D. degree in pathology from Ningxia Medical University, Yinchuan, China, in 2019. He is currently a Pathologist with the Department of Pathology, General Hospital, Ningxia Medical University. His current research interests include medical image processing and computational pathology.

XUEFEN ZHAO received the Ph.D. degree in mathematics and in applied mathematics from Ningxia University, Yinchuan, China, in 2016. She is currently an Associate Professor with the School of Information Engineering, Ningxia University. Her current research interests include deep learning and the application of artificial intelligence in fracture mechanics.

HONGJUAN GAO received the B.S. and M.S. degrees in software and theory from Ningxia University, Yinchuan, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer application and technology from Northwest University, Xi'an, China. Her research interests include graph and image processing, machine learning, and computer vision.

JIALONG SI was born in Shanxi, China, in 1999. He is currently pursuing the master's degree in computer technology with Ningxia University, Yinchuan, China. His main research interest includes medical image processing and analyzing.

CHUNHUI SHI was born in Shandong, China, in 1998. He is currently pursuing the master's degree in computer technology with Ningxia University, Yinchuan, China. His main research interest includes neural network model compression.