A Rotation-invariant Texture ViT for Fine-Grained Recognition of Esophageal Cancer Endoscopic Ultrasound Images

Tianyi Liu[1], Shuaishuai Zhuang[3], Jiacheng Nie[1], Geng Chen[1], Yusheng Guo[4*], Guangquan Zhou[2], Jean-Louis Coatrieux[5,6,7], and Yang Chen[1*]

[1] School of Computer Science and Engineering, Southeast University
{liutianyi_,niejiacheng,gchen,chenyang.list}@seu.edu.cn
[2] School of Biological Science & Medical Engineering, Southeast University
guangquan.zhou@seu.edu.cn
[3] The First Affiliated Hospital of Nanjing Medical University
shray1126@163.com
[4] Information Engineering, Xizang Minzu University
gys2777@163.com
[5] Centre de Recherche en Information Biomédicale Sino-Francais
[6] Inserm
[7] University of Rennes 1
jean-louis.coatrieux@univ-rennes1.fr

Abstract. Endoscopic Ultrasound (EUS) is advantageous in perceiving hierarchical changes in the esophageal tract wall for diagnosing submucosal tumors. However, the lesions often disrupt the structural integrity and fine-grained texture information of the esophageal layer, impeding the accurate diagnosis. Moreover, the lesions can appear in any radial position due to the characteristics of EUS imaging, further increasing the difficulty of diagnosis. In this study, we advance an automatic classification model by equipping the Vision Transformer (ViT), a state-of-the-art(SOTA) model, with a novel statistical rotation-invariant reinforcement mechanism dubbed SRRM. Mainly, we adaptively select crucial regions to avoid interference from irrelevant information in the image. Also, this model integrates histogram statistical features with rotation invariance into the self-attention mechanism, achieving bias-free capture of fine-grained information of lesions at arbitrary radial positions. Validated by in-house clinical and public data, SRRM-ViT has demonstrated remarkable performance improvements, suggesting our approach's efficacy and potential in EUS image classification.The source code is publicly available at: https://github.com/tianyiliu-lab/SRRM-ViT/.

Keywords: Fine-Grained Visual Classification (FGVC) · Endoscopic Ultrasound (EUS) · Rotation Invariant · Token Selection

* Corresponding authors: guangquan.zhou@seu.edu.cn, chenyang.list@seu.edu.cn.

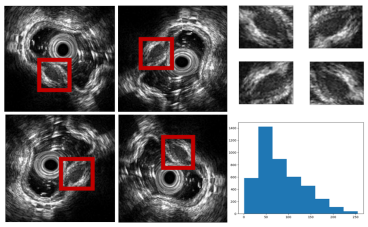


Fig. 1: The annular imaging characteristics of EUS result in a highly random spatial distribution of lesions. Firstly, we perform a rotation operation and then extract the lesion area from the detection image; following that, we apply histogram statistics to analyze this segmented region. The histogram analysis results showed that the histogram distribution of the lesion area remained unchanged.

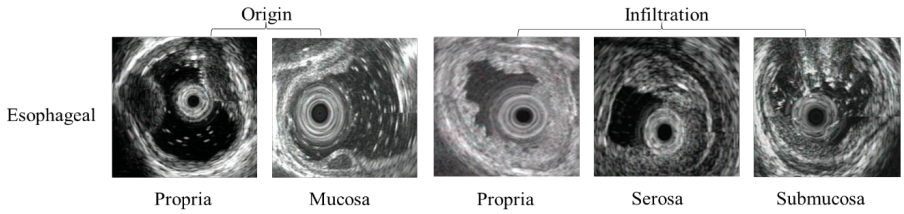


Fig. 2: Typical esophageal cancer presented in EUS of in-house dataset.

1 Introduction

Esophageal cancer is the eighth most common type of cancer worldwide and constitutes the sixth leading cause of cancer deaths [37]. Esophageal cancer detection methods include X-ray, CT, endoscopy, endoscopic ultrasound (EUS) technology, and tissue excision. Considering factors such as economy and security, EUS is more appropriate. Therefore, it is crucial to develop an accurate and robust EUS-based detection method in modern esophageal cancer research.

Due to the complex hierarchical structure of the esophagus, the types of lesions are diverse. As demonstrated in Fig.1 and Fig.2, the tumor originating from the muscular layer compresses the mucosal layer, and the infiltrative tumors directly invade both the mucosal and muscular layers, causing severe disruption to the structure. These characteristics lead to minor internal imaging differences within subclasses, contributing to a high misdiagnosis rate. Previous EUS detection of the digestive tract mainly focused on tumors originating within the gastrointestinal [54]. To our knowledge, our work is the first work to include infiltrative tumors in the classification task. Traditional classification models lack attention to internal detailed texture information, performing poorly on such complex problems. Therefore, a more fine-grained classification method is required for esophageal cancer ultrasound images. Current research methods in Fine-Grained Visual Classification (FGVC) can generally divided

into location-based and attention-based approaches. The localization-based approaches [4, 18, 45] achieve FGVC by directly annotating discriminative parts in the image. Previous work on esophageal cancer EUS grading based on this approach, where doctors segmented the lesion area first and then trained and predicted the segmented regions. However, pre-segmentation of regions is time-consuming and requires extensive clinical experience, which hinders the practical applicability of these methods. The attention-based methods [44, 53, 58] can detect discriminative regions in the image automatically through attention mechanisms. These approaches eliminate the need for handcrafted annotation of discriminative regions and have achieved encouraging results. However, enhancing rotation invariance pivot to the Esophageal Cancer EUS FGVC model has yet to be thoroughly explored in transformer-based approaches.

In this paper, we introduce Soft Histogram Texture Feature Reinforced Token Selection ViT(SRRM-ViT), an innovative FGVC technique for esophageal cancer diagnosis. To the best of our knowledge, this handcrafted feature integration mechanism is introduced for the first time, providing a novel approach to improve model performance and reinforce model robustness. In contrast to prior works [43, 55], where doctors segmented the lesion area first and then trained and predicted the segmented regions, our method avoids region pre-segmentation that is time-consuming and requires extensive clinical experience. As demonstrated in Fig.1, the local detailed texture information of the lesion and its surroundings plays a crucial role in the classification results; at the same time, the annular imaging characteristics of EUS result in a highly random spatial distribution of lesions within the images. Therefore, motivated by clinical experience, we first employed a self-attention [38] weight integration mechanism [16] for the EUS FGVC task in Esophageal Cancer. This method dynamically selecting crucial texture regions in the images to ensure the model makes decisions based on the distinctive texture features of different pathological categories. Moreover, considering CNN and ViT are translation invariant but sensitive to rotation [23,24], we further specifically focus on the issue of rotational invariance learning in EUS images to improve our method. Joshua Peeples et.al [28] have demonstrated that soft histogram features provide a degree of rotational invariance by learning the same local pixel distribution for each image, enhancing the performance of texture analysis. Inspired by this study, we advance a feature fusion mechanism integrating soft histogram features with rotation-invariant texture characteristics into the model, boosting its performance and mitigating rotation sensitivity. Experimental results on in-house data and publicly available dataset reveal outstanding performance of the proposed method, implying its applicability in various scenarios.

In summary, we make the following contributions:

- We employ a region importance discrimination model based on the ViT for EUS FGVC task in Esophageal Cancer, leading to a further improvement in performance.

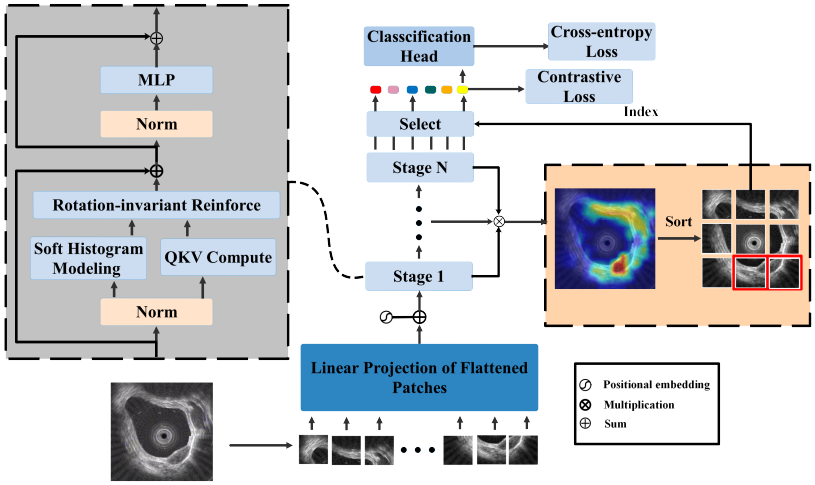


Fig. 3: Proposed architecture, including the Soft Histogram Texture Feature Modeling, Rotation-invariant Reinforcement, Token Selection Module, and the loss function.

- We propose the Soft Histogram Texture Feature Reinforce mechanism, integrating rotation invariant features into the key and value matrices to capture lesions in arbitrary spatial locations.
- For the first time, we use complete EUS images and include infiltrative tumors in the classification task. Compared to classical algorithms on both in-house and public data, our approach shows significant improvement.

2 Related Works

2.1 Endoscopic Ultrasound Image Classification

The detection rate of submucosal tumors (SMT) has been on the rise due to heightened public health awareness, the widespread popularity of endoscopic examinations, and the advancement and maturation of endoscopic ultrasound (EUS) technology [13, 36]. Recently, there have been some studies focusing on endoscopic ultrasound images of the digestive tract. Hangbin Zheng et.al [54] proposed a two-stage deep learning algorithm, which encompasses an attention-based network and self-supervised pre-training. Junke Wu et.al [43] proposed a multi-feature fusion classification method for adaptive EUS tumor images. They classified local lesions processed through principal component analysis, local binary pattern, and gray-level co-occurrence matrix using support vector machines. Hangbin Zheng et.al [55] introduce a Multi-Attribute Guided Contextual Attention Network (MAG-CA-Net) for interpretable SMT recognition in EUS. This network localizes abnormal areas using echo attributes and then determines tumor categories through contextual semantics. These tasks require doctors to pre-segment the lesion area and exclude infiltrative tumor categories.

2.2 Fine-Grained Visual Classification

Considerable efforts have been devoted to addressing fine-grained visual classification challenges. The primary obstacle in this domain lies in enhancing inter-class disparities and capturing intra-class similarities [49]. Currently, deep learning-based methods for fine-grained visual classification can be categorized into two main types. The first type is the localization method, which trains a detection network to accurately identify discriminative regions, followed by the fusion of these localized regions for fine-grained classification [8,14,21,49,52,57]. The second method is the feature encoding method, which aims to learn more informative features by calculating high-order information [20, 56] or exploring relationships between contrastive pairs [6, 7, 10, 29, 33, 35].

In recent years, ViT has demonstrated outstanding performance in general image classification [9], image retrieval [12], and semantic segmentation [58]. These achievements show that the inherent attention mechanism of the pure Transformer architecture can autonomously identify important image regions for recognition. As the first work investigating visual transformers in the FGVC domain [16], proposed replacing the input of the final transformer layer with key tokens, leading to improved results. Jun Wang et.al [40] introduced a new transformer-based framework, Feature Fusion Vision Transformer (FFVT), which aggregates key tokens from each transformer layer to capture local, low-level, and mid-level information. Wang, Q et.al [41] designed an attention aggregating transformer (AA-Trans) with a core attention aggregator (CAA) and an innovative information entropy selector (IES) to guide the network in acquiring discriminative parts of the image precisely. Sun, H et.al [34] introduced a structure information learning (SIL) module and a multi-level feature boosting (MFB) module, which incorporate spatial context information of key patches and utilize multi-level feature complementarity and inter-class contrastive learning to enhance feature robustness. These methods have not been tailored to enhance specific task characteristics, such as emphasizing rotation-invariant texture features in the fine-grained classification of esophageal cancer endoscopic ultrasound images.

2.3 Rotation-invariant Handcrafted Features

Combining neural and handcrafted features into a deep learning architecture can reinforce the model's learning ability for fine-grained texture features; these approaches have already been successful [5, 22, 27, 39, 42]. Additionally, the rotation invariance of handcrafted features can improve the model's sensitivity to rotation issues. However, handcrafted features can not be updated through backpropagation.

In recent years, much research have been conducted on the design task of soft handcrafted features, and significant progress has been made. Qiuze Yu et.al [50] proposed an improved nonlinear scale-invariant feature transform (SIFT). This framework-based algorithm combines spatial feature detection with local frequency-domain description to registering Synthetic Aperture Radar(SAR)

and optical images; these methods achieved remarkable accuracy improvements. Rohan Dubey et.al [11] proposed integrating the geometric structure of the Euclidean motion group SE(2) into convolutional networks via SE(2)-group convolution layers. This approach ensures translation and rotation equivariance, maintaining output invariance under a discrete set of rotations. Imran Riaz et.al [30] proposed a circular shift combination local binary pattern (CSC-LBP) to compute the local binary pattern within a 3×3 spatial window for each neighborhood pixel separately can obtain the more discriminative feature vector. Maxime W et.al [19] proposed a framework that encodes the geometric structure of the Euclidean motion group SE(2) into convolutional networks, achieving translation and rotation equivariance through SE(2)-group convolution layers. They ensured that the outputs remained invariant under a discrete set of rotations. Qi Xie et.al [46] explored an ameliorated Fourier series expansion for 2D filters and constructed a new equivariant convolution method; based on this, they proposed filter parametrization method, named F-Conv. In super-resolution image processing, the task outperforms classical convolution-based methods. Inspired by these works, we reinforce the ViT model with soft histogram features that can provide rotation-invariant texture information.

3 Methods

This section introduces SRRM-ViT in four parts: soft histogram feature modeling, rotation-invariant reinforcement, token selection and the loss function.

3.1 Overview

An overview of our method is shown in Fig.3. Firstly, we employ convolutional layers to model soft histogram features. Secondly, we construct the soft histogram rotation-invariant reinforcement mechanism, merging soft histogram features with self-attention features using the rotation-invariant reinforcement mechanism. Thirdly, we implement a module for sorting and selecting tokens based on attention weights, aggregating attention weights from each layer, and identifying tokens that capture intricate local lesion characteristics features. Finally, SRRM-ViT is trained with a hybrid approach combining contrastive loss and cross-entropy loss. This training strategy is designed to reduce the similarity between classification tokens of different labels while maximizing the similarity between classification tokens of samples with the same label.

3.2 Rotation-invariant Texture Self-attention

Soft Histogram Texture Feature Modeling. We utilize Radial Basis Functions (RBFs) [29] for histogram feature modeling. The smooth slope will create a path for gradients to flow backward through each histogram bin to the previous layer. In addition, RBFs exhibit robustness to slight variations in bin centers and widths compared to standard histogram operations. This is attributed to

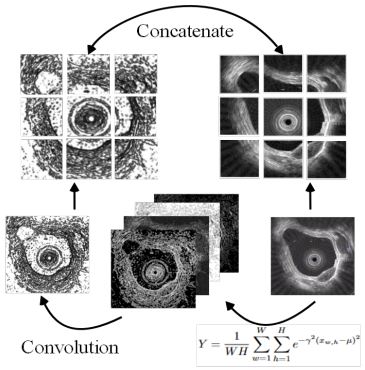


Fig. 4: Soft Histogram Texture Feature Modeling.

the soft bin allocation and smoothness of RBFs, allowing for a certain degree of error tolerance. Similar to previous studies [48, 51], but considering the low contrast of ultrasound images, we use a 3×3 convolutional kernel instead of a 1×1 kernel to capture more local context information, thus reducing the impact of image noise and extracting more texture features. We employ a $3 \times 3 \times D$ convolution to extract the feature maps from the input $x \in \mathbb{R}^{W \times H \times D}$, where d is the channel number of the input. The binning operation for histogram values is defined as follows:

$$Y_{rcnd} = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} e^{-\gamma_{nd}^2 (x_{r+w,c+h,d} - \mu_{nd})^2}, \tag{1}$$

where r and c denote the spatial dimensions of the histogram feature map, n represents the number of bins, and d represents the number of channels. The aggregation process of feature mapping is shown in Fig.4. The binning process begins by assigning each feature value to the nearest bin center (μ_{nd}). The center of each bin's feature is computed using a 3×3 convolution applied to each feature map, where the convolution kernel weights are set uniformly to 1, , and each bias acts as a trainable bin center. The width of features assigned to each bin (γ_{nd}) is determined by another 3×3 convolution applied to each feature map, where the biases in the convolution kernels are fixed at 0 and each weight is a trainable bin width parameter. B represents the number of bins. Finally, the feature map will be normalized by the exponential function e after passing through two special convolutional layers.

Rotation-invariant Reinforcement. To address the issue of the sensitivity of ViT to rotation, we fuse soft histogram feature map into weight of self-attention reinforce the attention of the ViT to rotation invariant features. Soft histogram reinforced self-attention can be viewed as a novel attention reinforcement learning method. Firstly, we use 1×1 convolution operation to aggregate all bins, generating a feature aggregation map with the same size as the original feature. Following that, the raw input and soft histogram feature aggregation map are

preliminarily divided into non-overlapped patches. The number of patches can be calculated as follows:

$$N = \frac{W \times H}{P^2}, \tag{2}$$

where N denotes the number of patches, and P represents the height and width of each patch. Patches are mapped into 1D vectors and encode the spatial positions and relationships of the patches after the previous step, calculated as follows:

$$\mathbf{z} = \left[x_p^1\mathbf{E}, x_p^2\mathbf{E}, \cdots, x_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \tag{3}$$

where N is the number of image patches, $\mathbf{E} \in \mathbb{R}^{(p^2 \times D) \times M}$ is the patch embedding projection, and $\mathbf{E}_{pos} \in \mathbb{R}^{N*M}$ denotes the position embedding, M represents the 1D vector dimension. Finally, we construct the reinforcement mechanism using original feature maps and soft histogram feature maps (Img1, Img2). For Img1, we compute the query, key, and value vectors, corresponding to $Q1$, $K1$, and $V1$. For Img2, we only compute the key and value vectors, corresponding to $K2$ and $V2$. Subsequently, we concatenate the key and value matrices from both images and compute the attention between the query of the target image and the combined key-value pairs as follows:

$$f\left(Q_1, K_{cat}, V_{cat}\right) = \text{softmax}\left(\frac{Q_1 K_{cat}^T}{\sqrt{d}}\right) V_{cat}, \tag{4}$$

where $K_{cat} = [K_1; K_2] \in \mathbb{R}^{(2N+2) \times d}$ and $V_{cat} = [V_1; V_2] \in \mathbb{R}^{(2N+2) \times d}$. We compute $N+1$ self-attention scores within itself and $N+1$ attention scores according to Eq.4. All the $2N+2$ attention scores are jointly normalized by the softmax function, thereby learning the attention scores for the target image Img1.

3.3 Token Selection Module

For the fine-grained classification of esophageal cancer EUS images, we need to identify subtle variations in the lesion areas. Benefiting from the patch-embedding mechanism of ViT, we can skip the region segmentation step and select tokens that are beneficial for classification. Due to the lack of token distinguishability in embeddings, the original attention weights may not accurately reflect the relative importance of input tokens, especially for higher layers of the model [1,31]. Therefore, we integrate the attention weights from the early layers. Specifically, we recursively apply matrix multiplication to the original attention weights of all layers. Due to the change in attention weight size caused by the rotation-invariant reinforcement mechanism, in contrast to previous work [16], the recursive process requires the addition of a transpose operation as:

$$a_{\text{final}} = \prod_{l=0}^{l-1} a_l' \quad a_l' = \begin{cases} a_l^T, l \bmod 2 \neq 0 \\ a_l\ , l \bmod 2 = 0. \end{cases} \tag{5}$$

As a_{final} captures information propagation from the input layer to embeddings in higher layers, this recursive process is preferred for selecting discriminative regions over single-layer raw attention weights a_{L-1}. We then identify the

indices corresponding to the maximum values $A_1, A_2, \cdots, A_K$ across K different attention heads in a_{final}. These indices correspond to the labels in z_{L-1}. Finally, we concatenate these selected labels with the classification labels to form the input sequence, denoted as:

$$\mathbf{z}_{\text{local}} = \left[z_{L-1}^0; z_{L-1}^{A_1}, z_{L-1}^{A_2}, \cdots, z_{L-1}^{A_K}\right]. \tag{6}$$

This operation preserves global information and enables the final Transformer layer to emphasize subtle differences between subclasses while disregarding less distinctive regions like background or common features.

3.4 Loss Function

We adopt the first token z_i of the Self-attention weight integration module for classification. The limitations of a basic cross-entropy loss become apparent in fine-grained classification tasks, where subtle distinctions between sub-categories may pose challenges for effective supervision. To address this, we utilize the contrastive loss ($\mathcal{L}_{con}$), which has been widely applied in FGVC tasks, to reduce the similarity between classification tokens associated with different labels while maximizing the similarity of classification tokens from samples sharing the same label y. To alleviate the influence of less challenging negative pairs, where samples from different classes show minimal similarity, we introduce a constant margin α. The $\mathcal{L}_{con}$, only considers negative pairs with a similarity greater than α, denoted as:

$$\begin{aligned}\mathcal{L}_{\text{con}} = & \frac{1}{N^2}\sum_{i}^{N}\left[\sum_{j:y_i=y_j}^{N}\left(1-\text{Cossim}\left(z_i,z_j\right)+\right.\right. \\ & \left.\left.\sum_{j:y_i\neq y_j}^{N}\max\left(\left(\text{Cossim}\left(z_i,z_j\right)-\alpha\right),0\right)\right]\right.\end{aligned} \tag{7}$$

$\mathcal{L}_{\text{cross}}\ (y, y')$ represents the cross-entropy loss between the predicted label y' and the ground-truth label y, denoted as:

$$\mathcal{L}_{\text{cross}}\ (y, y') = -\frac{1}{N}\sum_{i}^{N}\cdot\log\left(y_i'\right). \tag{8}$$

In summary, our model is trained with the combined loss function, consisting of the cross-entropy loss ($\mathcal{L}_{\text{cross}}$) and the contrastive loss ($\mathcal{L}_{\text{con}}$), expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{cross}}\ (y, y') + \mathcal{L}_{\text{con}}\ (z). \tag{9}$$

4 Experiments

4.1 Datasets

In-house Dataset. We utilized in-house data from the the First Affiliated Hospital of Nanjing Medical University, captured using Olympus Endoscopic Ultrasound from May 2011 to October 2020 and the class labels were provided by five

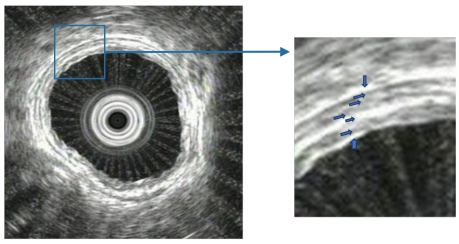


Fig. 5: Esophageal endoscopic ultrasound image hierarchical structure display.

Table 1: Distribution in the In-house Dataset.

In-house Dataset	Super-class	Origin		Infiltration		
	Sub-class	Propria	Mucosa	Propria	Serosa	Submucosa
	Training Number	118	294	375	64	472
	Testing Number	31	74	95	16	119

experienced doctors. We evaluate effectiveness of our method with classical methods and the recent SOTA FGVC methods on this dataset, the institutional ethics committee has approved this study. During esophageal endoscopic ultrasound, doctors use a water-filled balloon combined with intracaVitary water-filling technology to conduct the examination. Normal esophageal endoscopic ultrasound examination shows a seven-layer structure of the esophageal wall. The first to seventh layers from the inside to the outside are the hyperechoic mucosal epithelium, the hypoechoic mucosal lamina propria, the hyperechoic submucosal layer, the hypoechoic superficial muscularis propria, and the hyperechoic muscularis propria. The echogenic myenteric membrane, the hypoechoic deep layer of the muscularis propria, and the hyperechoic adventitial layer as shown in Fig.5. In this study, due to the thinning of the esophageal wall caused by the compression of the probe, only a hyperechoic layer was visible in layers 1-3 in most pictures.as detailed in Table1. There are 5 categories in total, including from muscularis propria, from muscularis mucosae, origin muscularis propria, origin adventitia, and origin submucosa. Fig.2 shows some sample images.

Brain Tumor MRI Dataset. Due to the lack of publicly available fine-grained ultrasound classification datasets and our intention to test the effectiveness of SRRM-ViT on different modalities of medical imaging data, we included a brain tumor MRI dataset [26] for comparison. The dataset comprises 7023 grayscale JPG images of human brain MRI, sourced from a combination of Figshare, SARTAJ, and BrH35 datasets, detailed in the supplementary material. It includes four classes of brain tumors: Glioma, Meningioma, No tumor, and Pituitary. Notably, images from the Glioma class in the SARTAJ dataset were initially misclassified and subsequently corrected during the dataset integration process. Images categorized as no tumor originated from the BrH35 dataset.

4.2 Training Details

We use 1324 cases for the training and 334 cases for the testing, and we implement 5-fold cross-validation for the training set. Common data augmentation techniques were applied, including horizontal flipping, vertical flipping, and random rotation. Data is strictly partitioned by patient, ensuring that data from the same patient does not appear concurrently in both training and test sets. The image inputs are resized to 224 × 224. Each fold has 500 epochs, initializing with intermediate weights from the official ViT-b16 model pre-trained on ImageNet21k. The batch size is set to 16. The learning rate is 0.03. All experiments are conducted using the PyTorch framework on an Nvidia 3090 GPU for training, evaluation, and testing.

Table 2: Comparing our method with the classic models and recent SOTA FGVC models on In-house Dataset(Unit: %, FLOPs unit: G).

Method	Baseline	Accuracy	Precision	Recall	F1-score	FLOPs
ResNet-18 [17]	Resnet-18	68.9	88.4	31.4	26.1	1.82
ResNet-50 [17]	Resnet-50	69.5	88.4	30.6	24.7	4.13
VGG-16 [32]	Vgg-16	65.6	87.2	29.3	22.6	16.9
ViT [9]	ViT-B 16	61.6	88.4	29.3	22.2	16.9
IELT [47]	ViT-B 16	66.7	88.5	34.9	27.5	19.3
FFVT [40]	ViT-B 16	73.8	91.3	33.6	28.9	16.4
Cross-X [25]	Resnet-50	74.2	91.2	36.1	31.1	4.13
AA-Trans [41]	ViT-B 16	74.3	91.3	33.0	28.4	25.7
SIM-Trans [34]	ViT-B 16	75.9	91.8	37.0	32.5	16.9
ViT+TS	ViT-B 16	76.3	90.5	39.7	34.7	17.0
ViT+TS+RR	ViT-B 16	77.4	**92.3**	35.3	31.3	32.5
ViT+TS+$\mathcal{L}_{con}$	ViT-B 16	77.9	91.0	**41.8**	36.7	17.0
SRRM-ViT	ViT-B 16	**78.7**	92.1	41.1	**36.8**	32.5

4.3 Results and Analysis

We compare with classic models and recent SOTA FGVC models on in-house data and use ViT as our baseline. We test our proposed Rotation-invariant Reinforcement(RR), Token Selection module(TS), and the loss function(L_{con}) to detect their improvement effect on the baseline. From Table2, we can see that the model performance with the SRRM-ViT is significantly improved over some classic models and SOTA models in recent years. The baseline, reinforced with the TS module, achieved a remarkable improvement in classification accuracy, reaching 14.7%. The RR module enables the fine-grained classification model to focus more on rotation-invariant texture features, leading to a further improvement of 1.1% in accuracy. Contrastive loss effectively increases the distance between representations of similar sub-categories and decreases it between

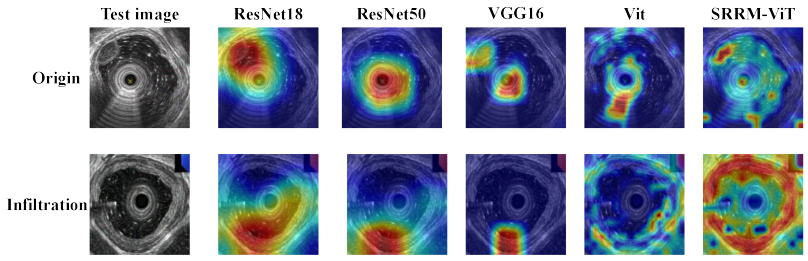


Fig. 6: The Grad-CAM heatmaps visualization of ResNet-18, ResNet-50, VGG16, ViT vs SRRM-ViT. The red-colored areas indicate high attention from the model.

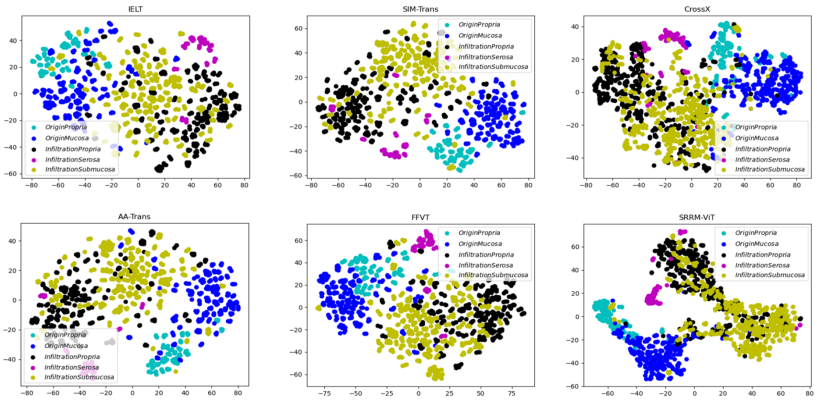
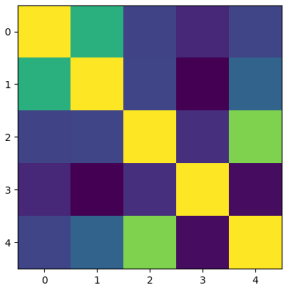
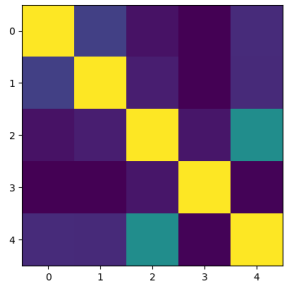


Fig. 7: The t-SNE visualization of IELT, SIM-Trans, CrossX, AA-Trans, FFVT vs SRRM-ViT implemented on test set.

identical categories, as observed in the comparison of the confusion matrix in Fig.8. The metric of the confusion matrix is cosine similarity. Viewing in color is recommended. The $\mathcal{L}_{\text{con}}$ further optimizes the model, ultimately achieving an accuracy of 78.7%. In medical image recognition tasks, model accuracy is the most critical evaluation metric for doctors. Compared to other models, our method also shows considerable advantages in other metrics. Although our approach introduces additional parameters, the FLOPs for testing a single image remain within an acceptable range. At the same time, we compared the sota model in recent years on public Brain Tumor MRI Dataset. As shown in Table 3, SRRM-ViT achieved an accuracy of 99.2%, which is 2.3% higher than the baseline ViT model(- stands for reference paper not provided). For a detailed analysis, please refer to the supplementary materials. This demonstrates the superiority of our proposed rotation-invariant texture enhancement mechanism across different modalities, proving not only its effectiveness but also the generalizability of SRRM-ViT.



(a) Confusion matrices of ViT.



(b) Confusion matrices of SRRM-ViT.

Fig. 8: Confusion matrices.

Fig.6 shows Grad-CAM heatmap visualization of ResNet18, ResNet50, VGG16, ViT and SRRM-ViT. ResNet18 and VGG16 recognize the origin of the lesion area but lack precision. Although CNN-based methods can focus on the lesion area in infiltrative data, the coverage is limited. It is evident that CNN-based methods emphasize local features while neglecting global features, leading to imprecise recognition of lesion areas. Although ViT exhibits global attention in infiltrative images, its accuracy is relatively low. Additionally, it is challenging for ViT to focus on lesions in origin-type images, aligning with its sensitivity to rotation. Our proposed method identifies rotation-invariant features and pays attention to the global information of the esophageal wall, which demonstrates the superiority of our proposed approach. It can be seen from Fig.7 that our method makes features of the same type more clustered than other methods. However, there is some confusion between the Infiltration Propria and Infiltration submucosa categories. We analyzed the original data and found that the irregularity of the infiltrative tumors caused the multi-layer structure of the esophagus to be destroyed at the same time, which affected the classification results. This limitation provides an inspiring foundation for our future work.

4.4 Ablation Study

We conduct ablation studies on SRRM-ViT to analyze how its variants affect the result of fine-grained classification tasks in esophageal cancer EUS images. All ablation studies are done on test sets. We evaluate the influence of the following designs: soft-histogram rotation-invariant reinforcement module, contrastive loss. The detailed analysis is as follows.

Influence of number of bin. The influence of various bin settings in Equation1 is shown in Table4. It is observed that a small number of bin can result in a reduction in rotation-invariant feature extraction capabilities, thereby reducing performance. Conversely, a high number of bin can increase model complexity and lead to model degradation. Empirically, we determine that 8 yields the best results in our experiments.

Table 3: Accuracy comparison with classic models and recent SOTA models on Brain Tumor MRI Dataset(Unit: %).

Method	Accuracy	Precision	Recall	F1-score
ConvAttenMixer [2]	97.9	98.6	95.3	96.7
Ensemble Learning [15]	98.0	-	-	-
CSA-MLP [3]	98.6	98.6	**98.5**	**98.3**
ViT [9]	96.9	98.6	73.0	72.2
SRRM-ViT	**99.2**	**99.6**	87.8	87.6

Table 4: Ablation study on number of bin and value of α (Unit: %).

Variate	Value	Accuracy	Precision	Recall	F1-score
Number of bin	4	75.6	91.2	35.7	31.3
	8	**78.7**	**92.1**	41.1	**36.8**
	12	77.9	91.5	**41.4**	36.8
	16	76.2	91.4	36.1	31.7
α	0.3	75.6	91.0	38.0	33.0
	0.5	75.3	91.0	35.4	30.6
	0.7	**78.7**	**92.1**	**41.1**	**36.8**
	0.8	77.4	91.8	40.5	36.3
	0.9	76.5	91.2	36.5	31.8

Influence of threshold for similarity α. The influence of different settings for the α in Equation7 is presented in Table4. It is observed that a small value of α can result in training signals being dominated by easy negatives, thereby reducing performance. Conversely, a high value of α can impede the model from learning sufficient information to increase the distances between hard negatives. Empirically, we determine that 0.7 yields the best results in our experiments.

5 Conclusion

In this paper, we first introduce the problem of the fine-grained classification task in esophageal cancer EUS images include infiltrative tumors class, provide a more comprehensive auxiliary diagnostic model for esophageal cancer detection. The rotation-invariant texture ViT we proposed not only solves the sensitivity problem of ViT models to lesion rotation, but also provides new ideas for hand-crafted feature enhancement methods. Finally, we achieved significant advances compared to classical models and recent SOTA models in both in-house data public data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62371121,T2225025), Jiangsu Provincial Key R&D Program(BE2022827) and the National Key R&D Program Project (2018YFA0704102).

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)
2. Alzahrani, S.M.: Convattenmixer: Brain tumor detection and type classification using convolutional mixer with external and self-attention mechanisms. Journal of King Saud University-Computer and Information Sciences **35**(10), 101810 (2023)
3. Arumugam, M., Thiyagarajan, A., Adhi, L., Alagar, S.: Crossover smell agent optimized multilayer perceptron for precise brain tumor classification on mri images. Expert Systems with Applications **238**, 121453 (2024)
4. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 955–962 (2013)
5. Cavalin, P., Oliveira, L.S.: A review of texture classification methods and databases. In: 2017 30th SIBGRAPI Conference on graphics, patterns and images tutorials (SIBGRAPI-T). pp. 1–8. IEEE (2017)
6. Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Wu, M., Guo, J., Song, Y.Z.: The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Transactions on Image Processing **29**, 4683–4695 (2020)
7. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5157–5166 (2019)
8. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6599–6608 (2019)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Pairwise confusion for fine-grained visual classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 70–86 (2018)
11. Dubey, R., Das, I.: Handwritten image detection using dcgan with sift and orb optical features. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON). pp. 1–6. IEEE (2023)
12. El-Nouby, A., Neverova, N., Laptev, I., Jégou, H.: Training vision transformers for image retrieval. arXiv preprint arXiv:2102.05644 (2021)
13. Fugazza, A., Fabbri, C., Di Mitri, R., Petrone, M.C., Colombo, M., Cugia, L., Amato, A., Forti, E., Binda, C., Maida, M., et al.: Eus-guided choledochoduodenostomy for malignant distal biliary obstruction after failed ercp: A retrospective nationwide analysis. Gastrointestinal Endoscopy **95**(5), 896–904 (2022)
14. Ge, W., Lin, X., Yu, Y.: Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3034–3043 (2019)
15. Hasan, M.A., Sarker, H., Roy, N.: Brain tumor detection using feature extraction and ensemble learning with a smart web application. In: 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). pp. 229–233. IEEE (2023)
16. He, J., Chen, J.N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C.: Transfg: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 852–860 (2022)

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked cnn for fine-grained visual categorization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1173–1182 (2016)
19. Lafarge, M.W., Bekkers, E.J., Pluim, J.P., Duits, R., Veta, M.: Roto-translation equivariant convolutional networks: Application to histopathology image analysis. Medical Image Analysis **68**, 101849 (2021)
20. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
21. Liu, C., Xie, H., Zha, Z.J., Ma, L., Yu, L., Zhang, Y.: Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11555–11562 (2020)
22. Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., Pietikäinen, M.: From bow to cnn: Two decades of texture representation for texture classification. International Journal of Computer Vision **127**, 74–109 (2019)
23. Liu, L., Chen, J., Zhao, G., Fieguth, P., Chen, X., Pietikäinen, M.: Texture classification in extreme scale variations using ganet. IEEE Transactions on Image Processing **28**(8), 3910–3922 (2019)
24. Liu, L., Fieguth, P., Wang, X., Pietikäinen, M., Hu, D.: Evaluation of lbp and deep texture descriptors with a new robustness benchmark. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 69–86. Springer (2016)
25. Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L.S., Li, J., Yang, J., Lim, S.N.: Cross-x learning for fine-grained visual categorization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8242–8251 (2019)
26. Nickparvar, M.: Brain tumor mri dataset. Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2645886.(Accessed on 3rd March) (2021)
27. Paul, R., Hawkins, S.H., Hall, L.O., Goldgof, D.B., Gillies, R.J.: Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic ct. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC). pp. 002570–002575. IEEE (2016)
28. Peeples, J., Xu, W., Zare, A.: Histogram layers for texture analysis. IEEE Transactions on Artificial Intelligence **3**(4), 541–552 (2021)
29. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
30. Riaz, I., Ali, A.N., Ibrahim, H.: Circular shift combination local binary pattern (csc-lbp) method for dorsal finger crease classification. Journal of King Saud University-Computer and Information Sciences **35**(8), 101667 (2023)
31. Serrano, S., Smith, N.A.: Is attention interpretable? arXiv preprint arXiv:1906.03731 (2019)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
33. Sun, G., Cholakkal, H., Khan, S., Khan, F., Shao, L.: Fine-grained recognition: Accounting for subtle differences between similar classes. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12047–12054 (2020)

34. Sun, H., He, X., Peng, Y.: Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5853–5861 (2022)
35. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the european conference on computer vision (ECCV). pp. 805–821 (2018)
36. Teoh, A.Y.B., Leung, C.H., Tam, P.T.H., Yeung, K.K.Y.A., Mok, R.C.Y., Chan, D.L., Chan, S.M., Yip, H.C., Chiu, P.W.Y., Ng, E.K.W.: Eus-guided gallbladder drainage versus laparoscopic cholecystectomy for acute cholecystitis: a propensity score analysis with 1-year follow-up data. Gastrointestinal endoscopy **93**(3), 577–583 (2021)
37. Uhlenhopp, D.J., Then, E.O., Sunkara, T., Gaduputi, V.: Epidemiology of esophageal cancer: update in global trends, etiology and risk factors. Clinical journal of gastroenterology **13**(6), 1010–1021 (2020)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., Madabhushi, A.: Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. Journal of Medical Imaging **1**(3), 034003–034003 (2014)
40. Wang, J., Yu, X., Gao, Y.: Feature fusion vision transformer for fine-grained visual categorization. arXiv preprint arXiv:2107.02341 (2021)
41. Wang, Q., Wang, J., Deng, H., Wu, X., Wang, Y., Hao, G.: Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification. Pattern Recognition **140**, 109547 (2023)
42. Wu, J., Lin, Z., Zha, H.: Multi-view common space learning for emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 464–471 (2016)
43. Wu, J., Wei, G., Wang, Y., Cai, J.: Multifeature fusion classification method for adaptive endoscopic ultrasonography tumor image. Ultrasound in Medicine & Biology **49**(4), 937–945 (2023)
44. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 842–850 (2015)
45. Xie, L., Tian, Q., Hong, R., Yan, S., Zhang, B.: Hierarchical part matching for fine-grained visual categorization. In: Proceedings of the IEEE international conference on computer vision. pp. 1641–1648 (2013)
46. Xie, Q., Zhao, Q., Xu, Z., Meng, D.: Fourier series expansion based filter parametrization for equivariant convolutions. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4537–4551 (2022)
47. Xu, Q., Wang, J., Jiang, B., Luo, B.: Fine-grained visual classification via internal ensemble learning transformer. IEEE Transactions on Multimedia (2023)
48. Xue, J., Zhang, H., Dana, K.: Deep texture manifold for ground terrain recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 558–567 (2018)
49. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 420–435 (2018)

50. Yu, Q., Ni, D., Jiang, Y., Yan, Y., An, J., Sun, T.: Universal sar and optical image registration via a novel sift framework based on nonlinear diffusion and a polar spatial-frequency descriptor. ISPRS Journal of Photogrammetry and Remote Sensing **171**, 1–17 (2021)
51. Zhang, H., Xue, J., Dana, K.: Deep ten: Texture encoding network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 708–717 (2017)
52. Zhang, Y., Wei, X.S., Wu, J., Cai, J., Lu, J., Nguyen, V.A., Do, M.N.: Weakly supervised fine-grained categorization with part-based image representation. IEEE Transactions on Image Processing **25**(4), 1713–1725 (2016)
53. Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S.: Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia **19**(6), 1245–1256 (2017)
54. Zheng, H., Bao, J., Dong, Z., Wan, X.: A data-efficient visual analytics method for human-centered diagnostic systems to endoscopic ultrasonography. Applied Intelligence **53**(24), 30822–30842 (2023)
55. Zheng, H., Dong, Z., Liu, T., Zheng, H., Wan, X., Bao, J.: Enhancing gastrointestinal submucosal tumor recognition in endoscopic ultrasonography: A novel multi-attribute guided contextual attention network. Expert Systems with Applications **242**, 122725 (2024)
56. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Learning deep bilinear transformation for fine-grained image representation. Advances in Neural Information Processing Systems **32** (2019)
57. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5012–5021 (2019)
58. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)