Patch-Based Multi-Level Attention Mechanism for Few-Shot Multi-Label Medical Image Classification

Mingyuan Li†, Yichuan Wang†, Junfeng Huang† , Erick Purwanto*, Ka Lok Man,
Xi'an Jiaotong-Liverpool University
Suzhou, China
e-mail: {Mingyuan.Li21, Yichuan.Wang21, JunFeng.Huang21}@student.xjtlu.edu.cn,
{Erick.Purwanto, Ka.Man}@xjtlu.edu.cn

***Abstract*—Few-shot learning stands as a prominent trend in the field of computer vision, with substantial applications in vision tasks such as image classification and semantic segmentation. It has gained popularity due to its potential to reduce the demand for computer resources and its ability to lessen dependence on large datasets. However, generating high-performance models becomes challenging since this approach must generalize only from a limited set of samples. This challenge is particularly evident in multi-label medical image classification, where overlapping labels and obscure characteristics within specific image regions impede the generalization capabilities of few-shot learning. This paper proposes a patch-based strategy with a multi-level attention mechanism. Our approach employs patch-based methods with multi-level attention to segment regions with overlapping information in images, thereby facilitating the extraction of crucial feature data. Experimental results reveal that the patch-based technique can help multiple models achieve greater classification performance across various datasets, demonstrating that the strategy effectively addresses the challenges inherent in multi-label classification.**

Keywords—few-shot learning; patch-based; multi-level attention; image classification; multi-label medical image

I. INTRODUCTION

In the field of medical image processing, deep learning algorithms have emerged as highly effective and promising methods for analyzing medical images [1]. However, their primary application remains in supervised learning, which poses a challenge for projects that lack a substantial number of annotated images since it necessitates a large amount of labeled data and substantial training costs across multiple iterations [2]. It becomes even more challenging for medical images because the acquisition and annotation of medical images are both time-consuming and expensive. Therefore, many researchers strive to develop models that exhibit superior performance without large labeled datasets. An approach addressing this challenge is few-shot learning [3].

Few-shot learning has achieved remarkable success across four categories: optimization-based, data augmentation-based, metric-based, and semantic-based [4]. Nevertheless, all of these few-shot approaches encounter challenges related to inadequate feature extraction [5]. These flaws will worsen when dealing with multi-label classification problems, resulting in a significant drop in the model's performance.

In the realm of image recognition and classification, the conventional single-label classification deals with the issue that an image can only be categorized under a single category. However, in real-world practical applications, images could have multiple semantic labels due to different semantic contexts across various regions [6]. Multimorbidity, the co-occurrence of distinct yet related diseases within a single image, is notably common in medical imaging [7]. Intricate connections exist between instances of the same disease. Moreover, the performance characteristics of a given disease can vary across datasets, thereby posing challenges to multi-label classification.

Since humans often prioritize visual feature identification over deep learning, Vaswani et al. [8] introduced the attention technique. This method enables the machine to concentrate its attention on specific regions. The attention mechanism finds diverse applications, including hyperspectral image recognition in remote sensing [9]. A subset of the attention mechanism is multi-level attention, which emphasizes attention across multiple levels. Through this approach, it becomes possible to identify the relevant features within the diverse levels of the image. Multi-level attention achieves tasks such as reusing and fusing multi-level features, preventing the blurring and loss of crucial shallow feature information, and enhancing deep features [10].

To this end, this paper proposed a patch-based strategy using a multi-level attention mechanism. In medical images with multi-label, overlapping label areas and subtle significant characteristics are common. Utilizing distinct patches can facilitate the identification of these crucial details. The multi-level attention mechanism encompasses pixel-level attention, intensifying focus on edges and textures. The patch-level attention prioritizes patches with more statistically significant features to receive higher attention. After combining the two levels, the subsequent preprocessing enhances the model's ability to emphasize the label components while disregarding other interfering information.

The aim of the research is to evaluate the performance of the patch-based preprocessing technique in multi-label classification tasks. The Experiment chapter will comprehensively describe the technique proposed in this research and introduce the datasets used in this study. The Experiment Results chapter will present experimental outcomes and evaluate the performance of the patch-based

† *Indicates equal contribution.*

* *Represents the corresponding author.*

strategy. Finally, the Conclusion chapter will summarize the findings and arguments presented in this paper, along with recommendations for addressing limitations and potential directions for future research.

II. Related Works

Multiple approaches are available for addressing multi-label classification challenges. Previous researchers have employed the strategy of dividing the problem into numerous binary splitting methods [11]. Contemporary multi-label classification issues analyze local and global information with greater depth. An example is the work by Ma [12], who introduced a novel method that can recognize and learn the interrelated prediction parameter matrix characteristics to focus on the high-order local information of many labels. However, the existing flaw in multi-label classification primarily manifests as inadequate classification outcomes due to the omission of certain targets.

Regarding attention, Yang [13] introduced an attention mechanism that relies on the highest value of the four-dimensional distribution probability of the image plate. This approach guarantees that each plate is assigned a specific lesion type. To better account for the unique distribution of features at multiple scales, a Multi-level Attention Capsule Network is proposed, allowing the acquisition of the degree spectrum features after multi-level wavelet decomposition [14]. Fully harnessing the potential of the multi-level attention mechanism can supply essential depth information to the model.

In the context of few-shot learning, the utilization of Visual Prompt Tuning proves highly beneficial [15]. According to Wang et al. [16], the approach involves initializing few-shot tuning using a pre-trained model from the Swin-transformer repository (pre-trained on ImageNet21K and fine-tuned on ImageNet1K). The subsequent results following this method exhibit a noteworthy enhancement in mAP compared to those not incorporating it.

In terms of patch-based approaches, as indicated by the study conducted by Rozario [17], the classification accuracy of the Bitumen dataset increased from 0.69 to 0.85, while the accuracy of the Asphalt dataset improved from 0.83 to 0.91. The mentioned dataset contains hyperspectral images, which have the property of decomposing the light in each pixel into multiple individual spectral bands. This feature exemplifies the substantial amount of information stored within each pixel of the images. Nevertheless, the improvement in accuracy indicates a significant potential for information extraction from hyperspectral images through the utilization of patch-based techniques. In another study by Hou [18], they developed a patch-level CNN with supervised decision fusion for the classification of Whole Slide Tissue Images. Due to the outstanding patch classification technique they devised, this unique patch-level classifier surpasses the performance of the image-level classifier. Furthermore, Zhang's [19] research concentrates on properties in each spatial domain and patch's channel, respectively, to gather crucial patch channel information and spatial information.

TABLE I. DETAILED PARAMETERS OF THE DATASET

Datasets Parameters	Chest	Colon	Endo
Size	512 x 512	1024 x 1024	1280 x 1024(Average)
Color Channel	RGB	RGB	RGB
Number of Training	2140	5654	1811
Number of Validation	2708	4355	2055
Number of Labels	19	2	4
Image Type	X-Ray Image	Cell Image	Colonoscopy Image
Target	Thoracic Diseases	Tumor Tissue	Lesion Types

Consequently, the feature information becomes more discriminative due to the introduced penalty system.

III. Experiment

A. Experiment Environment

The experiment was conducted using Google's Colab platform, utilizing a T4 GPU. The environment was configured with Python 3.10, Pytorch 1.8.0, Torchvision 0.9.0, Torchaudio 0.8.0, and Cudatoolkit 10.1. Additional required packages included: mmcls 0.25.0 (with its corresponding library openmim), scipy, scikit-learn, ftfy, regex, tqdm, and mmcv-full 1.6.0.

B. Dataset

To verify the effectiveness of the patch-based preprocessing with a multi-level attention mechanism in addressing few-shot classification for multi-label medical images, our team utilized the public dataset from the Foundation Model Prompting for Medical Image Classification Challenge 2023.

The dataset consists of three multi-label two-dimensional medical image subsets: Thoracic Disease Screening (ChestDR), Pathological Tumor Tissue Classification (Colon), and Lesion Detection in Colonoscopy Images (Endo). Specifically, the ChestDR dataset includes 4,848 frontal radiograph images (each 512×512) from 4,848 patients; the Colon dataset contains 10,009 large tissue patches (all with a uniform size of 1024×1024) from 396 patients; the Endo dataset consists of 3,865 images (with an average size of 1280×1024) from 80 patients [16]. Each subset is divided into a training and validation set with multi-labels. The ChestDR subset encompasses 19 common thoracic disease labels, the Endo subset includes 4 types of lesions, namely Ulcer, erosion, polyp, and tumor; and the Colon subset only has a tumor label. Further details can be found in Table 1.

C. Patch-Based Strategy

Our research focuses on an innovative preprocessing technique for enhancing few-shot classification efficiency in multi-label two-dimensional medical images. The methodology consists of segmenting imported medical

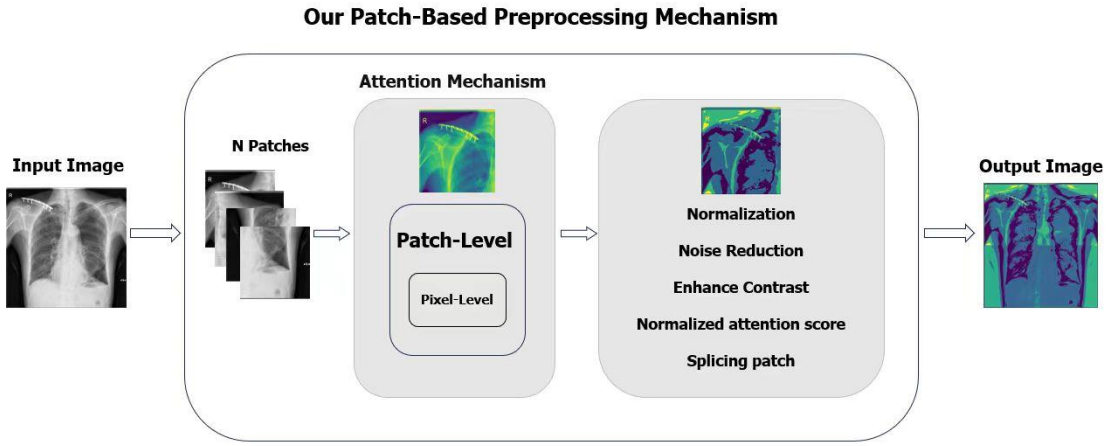


Figure 1. Patch-Based Strategy for Preprocessing with Multi-Level Attention Mechanism

images into patches, and each patch is subjected to image transformations such as standardization, contrast enhancement, and noise reduction. This preprocessing effectively addresses challenges in medical images like noise and intensity variations. For a detailed structure, please refer to Figure 1.

The proposed method addresses the multi-label issue in the experimental training set by dividing the original input image into small patches, corresponding to the number of labels in the dataset. These small patches collaborate with the multi-level attention mechanism to identify local regions of interest for further processing. After each patch undergoes the specific processing flow, the patches are stitched together to create the processed output image.

D. Multi-Level Attention Mechanism

Another pivotal facet of our methodology involves the implementation of a multi-level attention mechanism. This mechanism encompasses both pixel level and patch level components. At the pixel level, Sobel gradients are calculated along the x and y axes, allowing regions with pronounced intensity transitions to be highlighted by evaluating gradient magnitudes. At the pixel level attention directs subsequent processing stages towards relevant regions, thus enhancing feature extraction and classification accuracy. At the patch level, the attention mechanism bolsters the model's discriminative ability by concentrating on more extensive contextual information. The attention scores generated at this level are able to prioritize different patches in the image, ensuring that key features within various regions are recognized. Ultimately, the proposed method helps to aid in achieving more precise classification outcomes by incorporating diverse regions of interest.

For each patch, a comprehensive set of attention weights is created by combining the individual attention scores derived from both pixel and block levels. These amalgamated attention scores guide the linear enhancement of feature importance, optimizing the representation of each block within the image.

By leveraging the multi-level attention mechanism's power, the model can make more informed decisions during classification, ultimately resulting in improved few-shot classification performance, particularly in the complex field of multi-label 2D medical images. A comparison of the results after employing the patch-based strategy can be seen in Figures 2, 3, and 4.

IV. Experiment Results

A. Classification Performance Metric

We evaluate the quality of preprocessing by analyzing the results of multi-label classification performed on the preprocessed validation set. Specifically, the classification task aims to determine the classification status of the labels in the image.

Thus, to assess the performance of the preprocessing method, our team decided to use the metrics that reflect the accuracy of the model, namely mean Average Precision

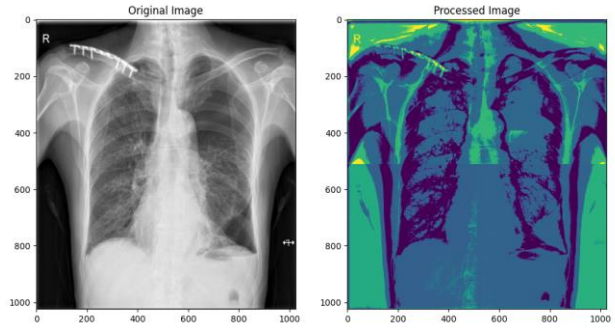


Figure 2. Original Chest image (left), highlighted regions signify prominent edges and important structures, such as the pulmonary lobe in the image (right)

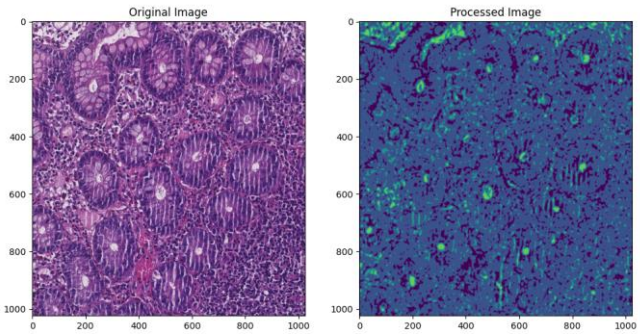


Figure 3. Original Colon image (left), highlighted regions with clear boundaries, such as cell nuclei (right)

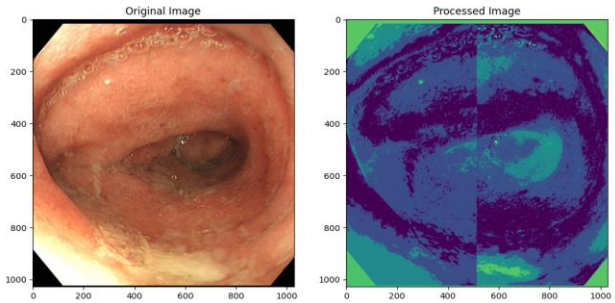


Figure 4. The original Endo image (left), highlighted regions suggestive of pathological changes, such as the ulcerous area in the lower portion of the image (right)

(mAP) and Accuracy (Acc), to reflect the performance of the preprocessing method indirectly. Additionally, to demonstrate the performance of the classifier models, we still used the Area Under Curve (AUC) metric. Although these metrics reflect the performance of the model, by comparing the model with the preprocessing part to the model without it, we can determine whether the preprocessing part improves the performance of the model and the value of the improvement.

In this experiment, the Accuracy metric, as defined in Eq. (1), is used to evaluate the performance of the model on the ChestDR and Endo datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

We use Mean Average Precision (mAP) to evaluate the performance of the models trained on the Colon dataset. The mAP formula is expressed in Eq. (2). In Eq. (2), N represents the total number of classes in the dataset, which is the count of distinct categories for evaluating the model's performance. APi denotes the Average Precision for the ith class.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{2}$$

Next, the calculation formula of AP is defined as Eq. (3). AP is a single-value summary of the Precision-Recall curve for a specific class, representing the model's ability to correctly classify instances of the i_{th} class while distinguishing them from instances of other classes. In Eq. (3), Rn stands for the recall rate at the n_{th} confidence threshold, and Pn represents the precision rate at the n_{th} confidence threshold. The formulas of Precision (P) and Recall (R) are provided in Eq. (4) and Eq. (5), respectively.

$$AP = \sum\nolimits_{n} (R_n - R_{n-1})P_n \tag{3}$$

$$Precision(P) = \frac{TP}{TP + FP} \tag{4}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{5}$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP + TN} \tag{6}$$

The final metric in our experimental results evaluation, Area Under Curve (AUC), refers to the area under the ROC curve. The ROC curve describes the relationship between the True Positive Rate (TPR, also known as the Recall rate) and False Positive Rate (FPR) as defined in Eq.(6) when the threshold is changed. On the ROC curve, the Y-axis represents TPR, and the X-axis represents FPR. In other words, AUC represents the area under the ROC curve and its value ranges between 0 and 1. AUC can be interpreted as the likelihood of randomly selecting a positive example and a negative example and assigning a higher score to the positive example than the negative one. Furthermore, the closer the AUC value is to 1, the more proficient the model is at correct classification.

B. Ablation Study on Three Datasets

To ensure that the model effectively considers both local details and global effects, all three datasets utilized an identical multi-level attention mechanism during the preprocessing phase.

For each experiment, we employed standard data augmentation techniques including contrast enhancement, noise reduction, etc. The ChestDR, Colon, and Endo datasets were used to evaluate the efficacy of the patch-based preprocessing approach, which led to the anticipated experimental outcomes. Across these datasets, the model's classification performance was enhanced by utilizing the patch-based preprocessing technique.

Furthermore, in the few-shot scenario, we conducted experiments on the three sets of medical images utilizing two exemplary networks: DenseNet and Swim-Transformer [16]. For the training of DenseNet, we fine-tuned the initial learning rate and utilized the SGD optimizer, respectively. The AdamW optimizer for the Swim-transformer model is accordingly fine-tuned. Using the MMClassification2 [20] framework, we trained these classification models over a span of 20 epochs, utilizing a batch size of 8 samples per iteration.

During the preprocessing stage, we implemented a multi-level attention mechanism involving two levels of attention: pixel-level and patch-level.

Pixel-level attention involves computing the Sobel gradient to ascertain the importance of each pixel. Pixels with larger magnitudes are considered more crucial, indicating their association with edges or regions of substantial change. Subsequently, the patch is multiplied with the gradient magnitude to emphasize important pixels and suppress unimportant ones. Its role is to focus on important local structures.

Patch-level attention leverages the mean and standard deviation of patches to calculate the attention score. Patches with high mean values and low standard deviation

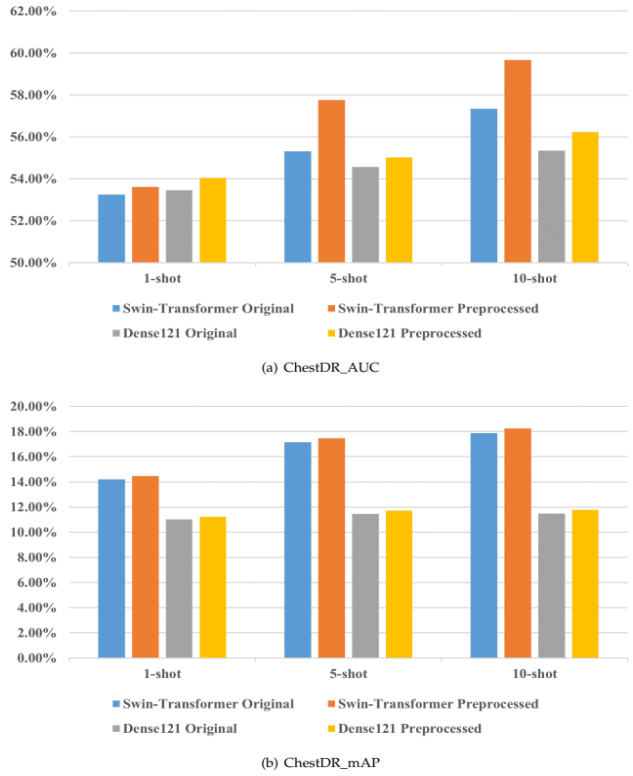


Figure 5. Evaluation results in 25 epochs on Chest: a slight improvement in mAP and AUC across all few-shot scenarios upon the utilization of patch-based

are identified as representative and stable, consequently receiving elevated scores. Following this, patches are weighted to emphasize these representative patches. The primary purpose is to pinpoint regions of heightened importance for classification.

At its essence, pixel-level attention assesses the value of individual pixels and integrates this information into the patch level. On the other hand, patch-level attention evaluates the overall effect of each patch and performs weighting. The comprehensive weight of each patch can be obtained by multiplying the two-level attention scores. The weighted patch is then employed to reconstruct the final image with the effect of attention focusing.

1) Experiment on ChestDR dataset

The ChestDR dataset comprises multiple chest X-ray images, accompanied by a CSV file containing labels. This dataset encompasses a total of 19 labels, and all images are standardized to a size of 1024×1024 pixels. To facilitate the equitable division of input images into patches roughly corresponding to the number of labels, we opted to divide the entire input image into 4×4 total of 16 patches. Generating this quantity of patches can minimize the training data requirement and shorten the training duration while maintaining proximity to the number of data labels.

Considering the common occurrence of noise in X-ray images, we implemented bilateral filtering as a noise reduction technique in this experiment. In this denoising method, the output value of each pixel is the weighted average of all pixels in its surrounding neighborhood. The weight is divided into two parts: spatial weight and gray value similarity weight. The pixels with close spatial distance and similar gray levels have significant weight. Bilateral filtering combines the spatial information and pixel value information of the image to smooth, which can preserve the edge details and avoid excessive blurring of the image.

To augment the distinguishability of the X-ray image and accentuate salient features, we apply contrast enhancement to the image. This enhancement makes some subtle features accentuate, and improves the visual quality of the image. This operation contributes to the model's ability to discern and learn crucial intricate patterns.

In the phase of few-shot training, we use the Visual Prompt Tuning (VPT) method to conduct the training based on the two types of Backbone mentioned above [16]. For the training of the DenseNet models, we utilize SGD optimizers with initial learning rates of 0.003 and 0.02 respectively. The AdamW optimizer uses an initial learning rate of 0.001 to optimize the Swim-transformer model. Additionally, we chose to set the epoch to 25 and the batch to 8.

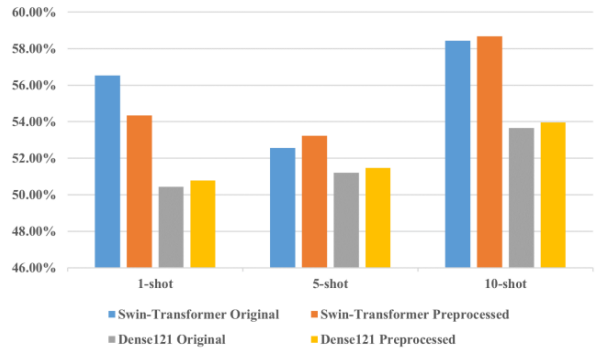
In terms of efficiency, Figure. 5's findings demonstrate a minor improvement from few-shot learning with patch-based methodology. Although the mAP under the Swin-transformer of 1-shot scenarios exhibits slight improvement, enhancements are evident in both the 5-shot and 10-shot scenarios. This divergence could potentially be attributed to the model acquiring incorrect key information during the 1-shot learning process. Notably, a significant observation is the substantial improvement in the 5-shot and 10-shot scenarios. This progress can be attributed to the model's increased emphasis on crucial information, which is highlighted through the preprocessing steps.

2) Experiment on Endo dataset

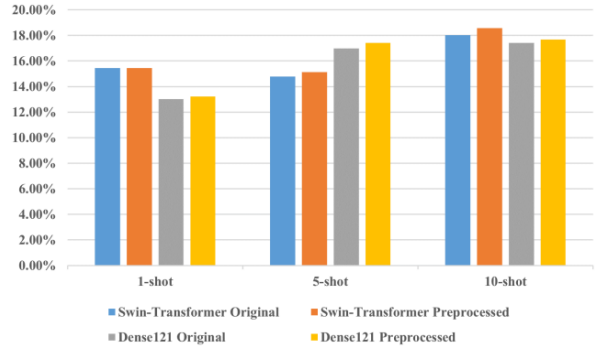
The images in the Endo dataset are genuine colonoscopy photos. The annotated CSV file within this dataset attributes only four labels to the images. This presents a notable reduction in label count compared to the 19 labels in the ChestDR dataset. Using the same quantity of patches as previously utilized could escalate the risk of overfitting, and significantly prolong training due to resource-intensive computations. Consequently, the number of patches in this training set was reduced from 16 to 4.

The environment in the Endo dataset is complex, with a significant portion of the pictures taken by colonoscopy including large areas of body fluids that reflect under the colonoscopy light. For images featuring extensive reflective zones, applying contrast enhancement may inadvertently amplify these reflective regions, potentially leading to counterproductive effects. Additionally, noise reduction processing was continued to minimize the impact of noise in the images.

We fine-tuned all baseline backbone models to accommodate the complex non-uniform illumination of the Endo. In terms of network structure, a dropout structure was added to deep networks like DenseNet, and a dropout structure was added to avoid overfitting when training on

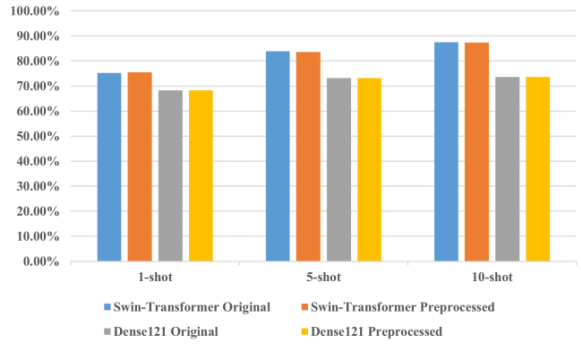


(a) Endo_AUC

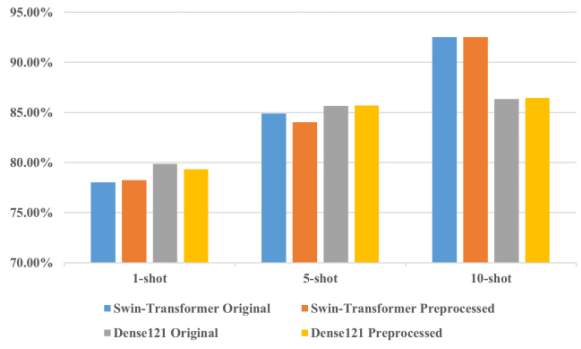


(b) Endo_mAP

Figure 6. Evaluation results in 20 epochs on Endo: a significant improvement when the models trained under 5-shot and 10-shot conditions using patch-based



(a) Colon_AUC



(b) Colon_Acc

Figure 7. Evaluation results in 20 epochs on Colon: the variance in Accuracy and AUC remains minimal and exhibits fluctuations across all few-shot scenarios upon the utilization of patch-based

preprocessed images. Simultaneously, adjustments were made to the baseline training strategy. In terms of learning rate, a strategy was employed. The learning rate increases linearly from a small value of 0.001 to 5 times that, and then decreases linearly again. Next, the overall batch size was dynamically adjusted, starting with a small batch size of 4 and gradually increasing it as the training proceeded, until reaching 16. To complement the above fine-tuning, the SGD optimizer was used, as it tends to yield better performance after prolonged training periods. In summary, these fine-tuning adjustments can assist the model in avoiding entrapment in local minima during the early stages, due to the complex lighting conditions in the images.

The classification model, trained on images preprocessed by our proposed patch-based strategy with the multi-level attention mechanism, exhibits a significant improvement in the score for the Endo. The experimental results are shown in Figure 6. Overall, the proposed preprocessing method exhibits strong performance at 5-shot and 10-shot levels, with only a slight improvement at the 1-shot level. Concurrently, the performance of this preprocessing method on the Swin-Transformer is markedly superior to that of other models. We believe that the patch-based strategy with the multi-level attention mechanism effectively enhances the region of interest and reduces interference in irrelevant areas, enabling the model to learn the main features more effectively.

3) Experiment on Colon dataset

The Colon dataset comprises microscopic photographs of colon cells. The limited label count for images within this dataset underlines our decision to preserve the existing patch count. Microscope images are considered to be more adaptable than other types of images, as they do not have reflective areas like those found in colonoscopy images. With the aim of accentuating target features, increasing classification separability, and enhancing model robustness, we continued to employ the same noise reduction, contrast enhancement, and patch normalization procedures as previously utilized.

Since the Colon dataset comprises only two labels, after fine-tuning the training parameters for SGD and AdamW optimizers, we decided to use 0.002 and 0.01, respectively. The same two backbones were trained using 25 epochs and 8 batches each. Analyzing the outcomes illustrated in Figure 7, it is evident that the accuracy after using the patch-based method fluctuates both above and below the original level. Given that the Colon dataset represents a binary classification problem, it may not be directly improved. Nevertheless, what is noteworthy is the absence of substantial disparity among different shots. In terms of the AUC, it can be seen that accuracy changes significantly with the increase in the number of shots.

V. Conclusion

In conclusion, while the few-shot model demonstrates strong performance in binary classification tasks, it encounters significant challenges when solving multi-label classification tasks. Hence, we have proposed a patch-based strategy that incorporates a multi-level attention mechanism. The results show that the patch-based strategy can effectively enhance the attention of labels by

combining pixel-level attention and patch-level attention, using the multi-level attention method.

Although improvements in performance metrics are noticeable for datasets processed using our proposed method, several shortcomings still remain. Firstly, after segmenting the original image into patches and employing the multi-level attention mechanism, fragmented pixel regions become evident at the intersections of patch regions, as illustrated in Figures 2, 3, and 4. This issue becomes pronounced when the number of patches is small (e.g., 2×2, 3×3), causing a complete object to be bisected into two disjointed parts, which can lead to training errors. Therefore, solving this problem may require increasing the number of patches and softening the transition of pixels at the borders of the patch areas. Secondly, a considerable computational overhead is present. As the number of patches increases, this overhead grows exponentially, primarily focusing on the calculations involved in the attention mechanism and image noise reduction. Thirdly, concerning the black edges of the image as illustrated in Figures 2 and 4, our attention mechanism algorithm will still allocate attention scores to these regions. Those scores have the potential to impede accurate assessments and consequently result in reduced precision.

To address the aforementioned problems, our future work will focus on finding a balance between performance and computational overhead, particularly when increasing the number of patches. This approach stems from our previous experiments, which show that preprocessing improved as the number of patches increased. Finding this balance point could significantly enhance the effectiveness of the preprocessing method used for few-shot multi-label classification in medical imaging. Furthermore, we will explore various computational methods to soften pixels at patch boundaries, aiming to avoid fragmentation in the image content caused by the original Patch-Based Strategy, which could negatively influence the model's training results. Simultaneously, we will continue to refine the attention algorithm to enable it to recognize black edge regions within patches and disregard them, thereby mitigating the impact of these edges.

ACKNOWLEDGMENT

This work is partially supported by the XJTLU AI University Research Centre and Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU. Also, it is partially funded by the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004) as well as funding: SURF-2023-0123, XJTLU-RDF-22-01-062, XJTLU-REF-21-01-002 and XJTLU Key Program Special Fund (KSF-A-17).

REFERENCES

[1] M. Ayyannan, M. A, S. D, N. T, T. M and N. S. Kumar, "Medical Image Classification using Deep Learning Techniques: A Review," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1327-1332, doi: 10.1109/ICEARS56392.2023.10084948.

[2] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1199-1208, doi: 10.1109/CVPR.2018.00131.

[3] C. -S. Cheng, H. -C. Shao and C. -W. Lin, "Task-Aware Few-Shot Visual Classification with Improved Self-Supervised Metric Learning," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 3531-3535, doi: 10.1109/ICIP46576.2022.9897878.

[4] J. Song, Z. Zhu, B. Li and S. Ni, "Few-shot Learning based on Multi-Attention and Prototype Correction," 2022 8th International Symposium on System Security, Safety, and Reliability (ISSSR), Chongqing, China, 2022, pp. 83-84, doi: 10.1109/ISSSR56778.2022.00020.

[5] Y. -X. Wang, R. Girshick, M. Hebert and B. Hariharan, "Low-Shot Learning from Imaginary Data," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7278-7286, doi: 10.1109/CVPR.2018.00760.

[6] C. Shen, M. Yi and M. Fu, "An improved multi-label classification algorithm based on YOLOV5s," 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, 2023, pp. 674-678, doi: 10.1109/NNICE58320.2023.10105674.

[7] Y. Xing, B. J. Meyer, M. Harandi, T. Drummond and Z. Ge, "Multimorbidity Content-Based Medical Image Retrieval and Disease Recognition Using Multi-Label Proxy Metric Learning," in IEEE Access, vol. 11, pp. 50165-50179, 2023, doi: 10.1109/ACCESS.2023.3278376.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need", Advances in neural information processing systems, pp. 5998-6008, 2017.

[9] Derbashi and E. Aptoula, "Scale-Spectral-Spatial Attention Network for Hyperspectral Image Classification," 2022 30th Signal Processing and Communications Applications Conference (SIU), Safranbolu, Turkey, 2022, pp. 1-4, doi: 10.1109/SIU55565.2022.9864719.

[10] X. Liu, L. Yang, X. Ma and H. Kuang, "Skin Disease Classification Based on Multi-level Feature Fusion and Attention Mechanism," 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2023, pp. 1500-1504, doi: 10.1109/ICIBA56860.2023.10165024.

[11] L. Jia, J. Fan, D. Sun, Q. Gao and Y. Lu, "Research on multi-label classification problems based on neural networks and label correlation," 2022 41st Chinese Control Conference (CCC), Hefei, China, 2022, pp. 7298-7302, doi: 10.23919/CCC55666.2022.9902377.

[12] Xiaoke Ma, Shiyin Tan, Xianghua Xie, Xiaoxiong Zhong, Jingjing Deng, Joint multi-label learning and feature extraction for temporal link prediction, Pattern Recognition, Volume 121, 2022, 108216, ISSN 0031-3203.

[13] Z. Xia, H. Hu, W. Li, Q. Jiang, C. Zhu and Z. Zou, "A Framework for Identifying Diabetic Retinopathy Based on patch attention and lesion location," 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 2023, pp. 1-8, doi: 10.1109/IJCNN54540.2023.10191557.

[14] Z. Tao, T. Wei and J. Li, "Wavelet Multi-Level Attention Capsule Network for Texture Classification," in IEEE Signal Processing Letters, vol. 28, pp. 1215-1219, 2021, doi: 10.1109/LSP.2021.3088052.

[15] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim,"Visual prompt tuning," in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Ciss´e, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 709–727.

[16] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, Q. Duan, J. Zhao, K. Li, Y. Qiao, and S. Zhang, “Medfmc: A realworld dataset and benchmark for foundation model daptation in medical image classification,” 2023.

[17] P. F. Rozario et al., "Deep Learning Patch-Based Approach for Hyperspectral Image Classification," 2023 IEEE International Conference on Electro Information Technology (eIT), Romeoville,

IL, USA, 2023, pp. 458-463, doi: 10.1109/eIT57321.2023.10187387.

[18] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis and J. H. Saltz, "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2424-2433, doi: 10.1109/CVPR.2016.266.

[19] X. Zhang, A. Deng, Y. Wang and Z. Wang, "Image Classification Algorithm combining Local Feature and Attention," 2022 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 2022, pp. 1226-1231, doi: 10.1109/IAECST57965.2022.10062079.

[20] Mmclassification. https://github.com/open-mmlab/mmclassificatio (accessed Aug. 20)