JTE HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Early Exit Based on Deep Learning Model for Polyp Colonoscopy Image Classification

Hoang Long Nguyen*, Minh-Vu Phan, The-Anh Pham
Hong Duc University, Vietnam

*Corresponding author. *Email: nguyenhoanglong@hdu.edu.vn*

ARTICLE INFO

Received: 15/11/2024
Revised: 08/12/2024
Accepted: 19/12/2024
Published: 28/08/2025

KEYWORDS

Early Exit;
Deep Learning;
Polyp;
Classification;
Computational efficiency.

ABSTRACT

Early exit is a widely adopted approach to reduce the inference time of deep learning models. By introducing side-branch classifiers into the main backbone network, this approach allows test samples to be predicted and exit the network early when high confidence is achieved. While the early exit mechanism has been extensively explored in various computer vision applications, its use in medical imaging remains relatively underexplored. In this study, we propose to design a lightweight early exit branch for polyp colonoscopy image classification with a combination of Convolutional Block Attention Module (CBAM) and Fully Connected Layer (FC). These branches are embedded into a deep learning backbone to leverage intermediate features for early predictions. Extensive experiments on the Kvasir polyp dataset demonstrate that our method achieves a favorable trade-off between accuracy and computational efficiency, showcasing its potential of lightweight early exit mechanisms to improve the efficiency of deep learning systems in medical image analysis, paving the way for faster and more resource-efficient diagnostic tools.

Doi: https://doi.org/10.54644/jte.2025.1721

1. Introduction

Over the past decade, with the advances of hardware devices and the availability of data, large and complex deep learning models have proposed and demonstrated effectiveness in computer vision tasks. Although scaling up deep learning models offers their capabilities, it exists some drawbacks such as carbon emissions and time-consuming training and inference. Furthermore, these models meet many practical challenges in deployment on real-life applications due to huge computational cost. Motivated by these challenges, several studies have been introduced to minimizing computational cost while improving efficiency of deep learning models including model quantization [1] [2], pruning [3] [4], knowledge distillation [5] [6], and early exit [7] [8].

Among these approaches, early exit has been denoted its efficiency and flexibility during inference stage. Early exit is a technique that applying several exit classifiers into the original deep learning networks, allowing samples to be predicted and stop the network with high confidence. Early exiting offers many aspects to improve computational efficiency such as feature reuse [9], training strategy [10] [11] and termination policies [12] [13]. However, the number of research of early exiting for medical image is still limit. Thus, in this paper, we propose to apply early exit side-branch for polyp colonoscopy image classification.

In this paper, we propose two key contributions. First, we propose a lightweight and effective side-branch classifier by integrating the Convolutional Block Attention Module (CBAM) with Fully Connected (FC) layers, enhancing the capability of intermediate predictions. Second, we evaluate the performance of a baseline model with the incorporation of Early Exit Branches for polyp classification. Through extensive experiments, we demonstrate that this approach not only improves classification accuracy but also significantly enhances computational efficiency. These contributions highlight the potential of our method in developing efficient and accurate solutions for medical image classification tasks.

2. Related works

Early exit. Teerapittayanon et al. [7] proposed BranchyNet to employ early exit classifiers to a deep learning model for image classification. BranchyNet deployed the entropy of prediction score and heuristic thresholds for early exit strategy. MSDNet was proposed by Huang et al. [14] further enhanced performance of early exit by using maximum prediction score instead of the entropy for measuring the exit confidence. Phuong and Lampert [11] employed a new training strategy based on knowledge distillation to improve earlier exits efficiency. Hu et al. [15] used data augmentation approach for optimizing model accuracy, robustness and efficiency.

Medical image classification. Convolutional Neural Networks (CNNs) have emerged as the dominant architecture, enabling remarkable performance improvements across various applications including image classification, object detection/segmentation. The introduction of ResNet-based models [16] [17] has significantly enhanced classification accuracy. In the field of medical image processing, Srinivas et al. [18] applied transfer learning approach based on CNN architectures such as VGG-16, ResNet-50, Inception-V3 to achieve competitive results for brain tumor classification task. Z Liu et al. [19] proposed a novel multi-scale convolutional neural to overcome the limitations of existing methods, which often fail to capture both the fine details and the overall structure of images effectively. A variant of ResNet, called ResGANet, was proposed in [20] stacking group attention blocks to capture both channel and spatial information in medical images. However, these models contain drawbacks in exploring contextual information, while global semantic features are also crucial in high-resolution medical images.

Inspired by the success of attention mechanism in Transformer [21] for natural language processing (NLP), significant efforts were proposed to apply Transformer architecture for computer vision. ViT [22] emerged as a successful approach in solving the drawback of capturing long-range dependencies in CNNs. ViT and its variants were widely used in computer vision tasks such as image classification, object detection, semantic segmentation. Transformer-based architecture have further applied into medical image processing tasks. TransFuse [23] was proposed to combine CNN and Transformer architectures in medical image segmentation. Dai et al. [24] introduced the first multi-modal medical image classification framework based on the advantages of CNN and Transformer. HiFuse [25] employed global and local feature blocks relying on Transformer architecture, while hierarchical feature fusion block (HFF block) was proposed to effectively fuse global and local features. Motivated by the advantages of HiFuse, in this study, we treat HiFuse as our baseline model to apply early exit side-branch classifier for enhancing accuracy and reducing inference cost.

3. Proposed method

3.1. Baseline model

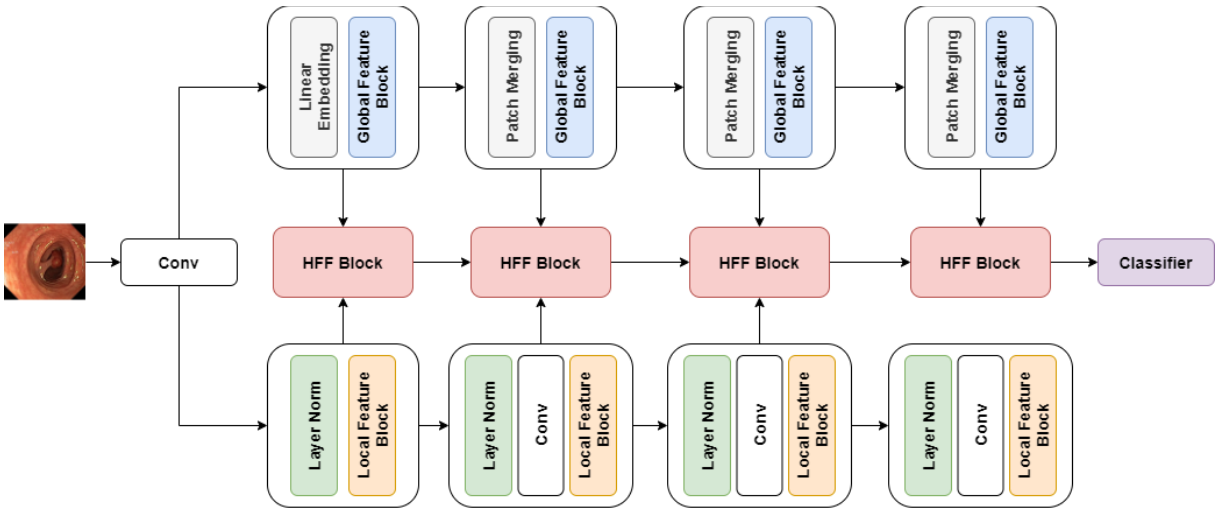

Figure 1. *Overview of HiFuse model. Reformed from* [25].

HiFuse [25] has demonstrated its effectiveness in medical image classification tasks, shown in Fig 1. Therefore, we decided to apply HiFuse as our baseline model for early exit strategy. Its approach focuses on capturing local spatial features and global semantic representations across multiple scales.

At each global feature extraction stage, after HiFuse applied Windows Multi-head Self-Attention (WMSA) in extracting global features. WMSA divided feature maps into smaller windows with the size of $M \times M$. Then, self-attention will be computed in each window and FFN has been replaced by GELU activation function to reduce computational cost. Shift WMSA (SWMSA) is also introduced in global feature block. Mathematically, this can be depicted as in Eqs 1, 2:

$$g_i = Conv_{1\times1}\left(WMSA\left(LN(G_{i-1})\right)\right) + G_{i-1} \tag{1}$$

$$G_i = Conv_{1\times1}\left(SWMSA\left(LN(g_i)\right)\right) + g_i \tag{2}$$

where F_i and f_i represent for the output of the SWMSA and WMSA block, respectively. LN denotes LayerNorm operation. $Conv_{1\times1}$ illustrates convolutional layer with the kernel size of 1×1.

Local feature extraction branch leverages the separability of convolutions by using 3×3 depthwise convolution leading to reduce complexity of the convolution as shown in Eq 3.

$$L_i = Conv_{1\times1}\left(LN(Conv_{3\times3}(L_{i-1}))\right) + L_{i-1} \tag{3}$$

where L_i serves as the output of the local feature block. $Conv_{3\times3}$ describes depthwise convolutional layer with the kernel size of 3×3.

These feature are further fed into hierarchical feature fusion block (HFF block) to effectively adapt and fuse. HFF leverages the capability of channel and spatial attention mechanisms. Specifically, the channel attention (CA) mechanism explores the interdependencies between channel maps to enhance the feature representation while the spatial attention (SA) mechanism improves essential regions selectively, detailed as in Eqs 4-10:

$$SA(x) = Sigmoid(Conv\left(Concat\left(AvgPool(x), MaxPool(x)\right)\right) \tag{4}$$

$$CA(x) = Sigmoid(MLP\left(AvgPool(x)\right) + MLP(MaxPool(x))) \tag{5}$$

$$L'_i = SA(L_i) \otimes L_i \tag{6}$$

$$G'_i = CA(G_i) \otimes G_i \tag{7}$$

$$F_i = AvgPool\left(Conv(F_{i-1})\right) \tag{8}$$

$$F'_i = Conv(Concat(L_i, G_i)) \tag{9}$$

$$F_{out} = IRMLP(LN(Concat(L'_i, F'_i, G'_i))) + F'_i \tag{10}$$

where $\otimes$ denotes element-wise multiplication operator. F_{i-1} represents for the output feature from previous HFF block. L'_i and G'_i are the output of spatical and channel attention mechanisms, respectively. IRMLP illustrates Inverted Residual MLP to generate semantic feature maps.

3.2. Early exit side-branch classifier

Though HiFuse achieved significant results in medical image classification, its computational cost is quite high. Thus, we propose to apply early exit approach into HiFuse model, called EE-HiFuse, depicted in Fig 2.

JTE
HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

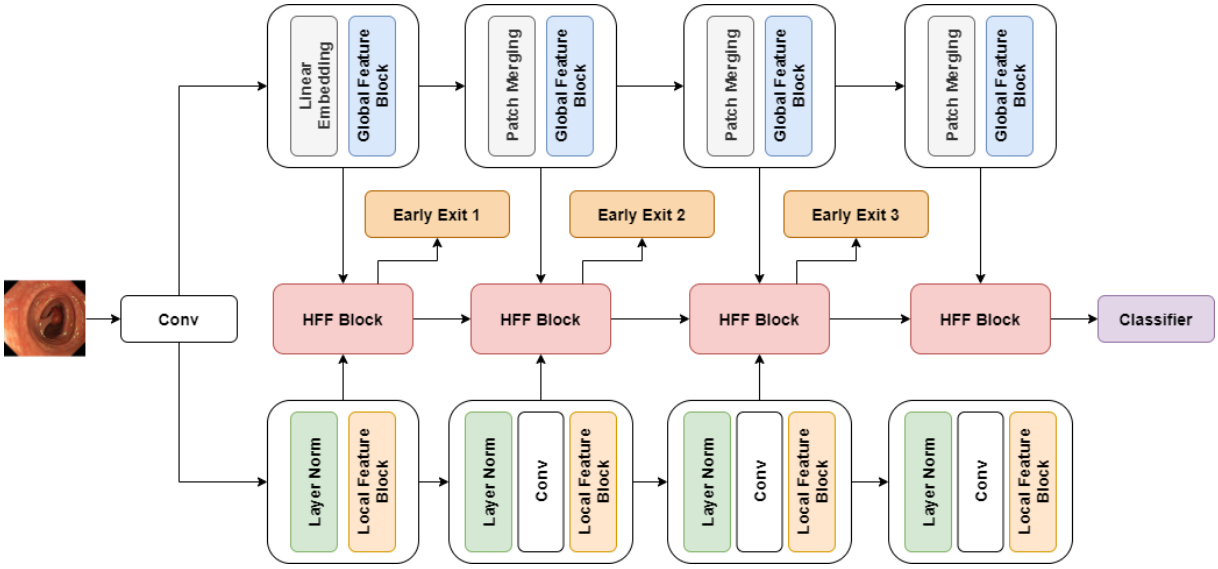

Figure 2. *Overview of our proposed EE-HiFuse model.*

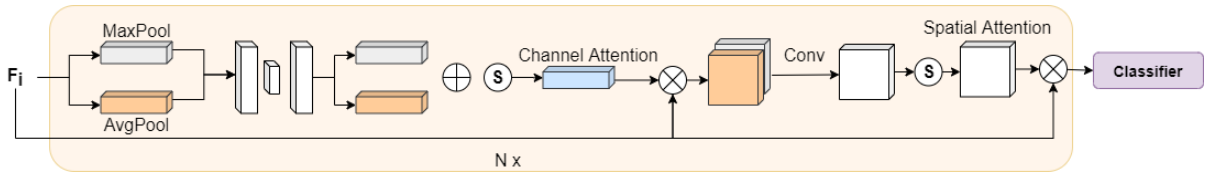

Figure 3. *Overview of our early exit architecture.*

The purpose of early exit side-branch classifier is to enable network to predict "easy" samples at earlier stages, eliminating the need to process the entire network. Assume that F_1, F_2, F_3, F_4 with $F_i \in R^{H' \times W' \times C}$ correspond the output feature maps from HFF blocks, where H' and W' denote the height and width of feature map F_i. To effectively capture important features from HFF blocks, we leverage the use of Convolutional Block Attention Mechanism (CBAM) [26]. The CBAM module contains two components: spatial attention and channel attention. Both spatial attention and channel attention modules in CBAM and HFF are the same. However, instead of processing feature simultaneously, CBAM firstly applied channel attention (CA) to focus on essential features over different channels, then spatial attention (SA) is employed to emphasize the importance of different locations. The output of CBAM can be described in Eqs 11, 12 as follow:

$$E_i = CA(F_i) \otimes F_i \tag{11}$$

$$E'_i = SA(E_i) \otimes E_i \tag{12}$$

Fig 3 illustrates our proposed early exit approach in baseline model. The multi-exit classifier branches are strategically integrated into the backbone, with each branch positioned between its stages. These branches enable early predictions by leveraging intermediate feature representations, thereby facilitating efficient inference while maintaining accuracy. Specifically, the backbone network generates multi-level features f_i, i ∈ {1,2,3,4}, which are processed through their respective exit points. This design ensures enhanced feature refinement and improves classification performance at each exit point.

Once side branches added, the network returns M classification outputs as in Eq 13:

$$[y_1, \dots, y_M] = [f_1(g_1(x), \theta_1), \dots, f_M(g_M(x), \theta_M)] \tag{13}$$

where x is the input image, y_i denotes the i^{th} soft-max output. f_i corresponds the i^{th} side branch classifier with the parameter θ_i.

During the inference phase, the classification output y_i is computed gradually from 1 to M. In this study, we apply a simple threshold for early exit t, in case, $\max(y_i) < t$, we decide to compute the next

y_{i+1} at the $i+1^{th}$ stage, otherwise, the sample will be processed at the next exit point in the i^{th} stage. In this study, we adopted the approach outlined in [7] to identify the threshold t that selects a configuration satisfying the specified constraints.

3.3. Loss function

In classification task, the soft-max cross entropy loss is usually applied for all the outputs. Let y denote the ground-truth vector, x is the input sample, $\hat{y}_i$ illustrates the output of the i^{th} side branch classifier. The loss function can be written as:

$$L_{total} = \sum_{i=1}^{M} CE(y, \hat{y}_i) \tag{14}$$

4. Results and Discussion

4.1. Dataset

Our proposed method is evaluated on a widely used dataset for polyp classification: Kvasir [27]. The Kvasir dataset contains 4000 colonoscopy images, annotated and verified by experienced doctors with different resolution from 720×576 up to 1920×1072 pixels. It includes 8 classes that display anatomical landmarks, pathological findings, and endoscopic procedures. Fig 3 shows several examples from the Kvasir dataset.

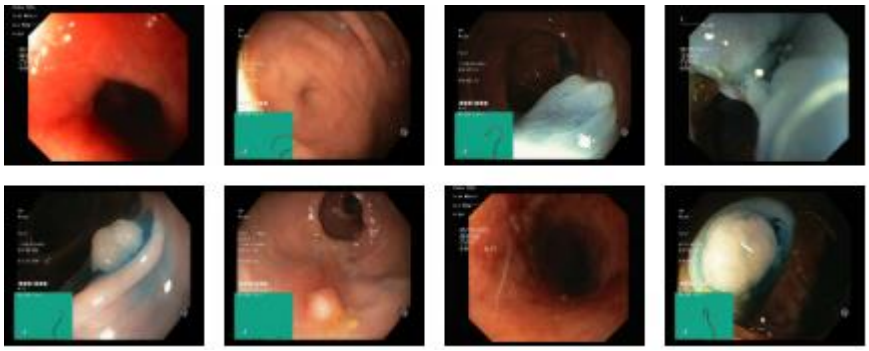

Figure 4. *Several examples extracted from the Kvasir dataset.*

4.2. Metrics and implementation

Metrics. In this study, we employ accuracy as our evaluation metrics. This metrics are based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Each metric will be detail below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

In addition, we compute the average inference time, number of parameters of each configuration and GFLOPs to demonstrate the trade-off between efficiency and computational cost of our proposed method.

Implementation. In our experiments, we split the Kvasir dataset into training and validation datasets with a ratio of 8:2. All experiments are based on Pytorch framework on the following hardware configurations:

Table 1. *Configurations of hardware devices*

Device	Name
GPU	GeForce RTX 2080Ti, 12GB
CPU	Intel® Xeon(R) W-2133

In addition, in the testing phase, we deploy the model on an only CPU configuration. We set the number of training parameters as follow:

Table 2. *Configurations of hyper-parameters*

Configuration	Detail
Number of epochs	100
Batch size	64
Image size	224×224
Optimizer	AdamW
Initial learning rate	$1e-4$
Number of CBAM blocks in EE	3

4.3. Results

As mentioned in the previous section, we evaluate our method against the original state-of-the-art method. The performance comparison results are displayed in Table 1.

Table 3. *Performance comparison between different early exits*

Model	Accuracy	Average Inference Time (seconds)	Number of parameters (M)	GFLOPs
HiFuse (Baseline)	86.6%	0.1865	121.614	18.142
EE-HiFuse (Early Exit 1)	82.8%	**0.0372**	**1.357**	**4.241**
EE-HiFuse (Early Exit 2)	**88.1%**	0.0629	6.938	8.648
EE-HiFuse (Early Exit 3)	87.1%	0.1269	32.798	13.747
EE-HiFuse (Final Classifier)	**88.1%**	0.1920	121.695	18.144

The performance detailed in Table 1 shows that our proposed method outperforms the original HiFuse model on the Kvasir classification dataset on accuracy metric. The baseline HiFuse model achieves an accuracy of 86.6%, with an average inference time of 0.1865 seconds, a substantial number of parameters with 121.614 million, and 18.142 GFLOPs, indicating a relatively high computational demand. Applying early exit strategy significantly reduces inference time and the complexity of the model as well.

In the first early exit configuration, the EE-HiFuse demonstrates the most efficient setup in terms of average inference time with 0.0372 seconds and the number of parameters with 1.357 million, as well as the lowest GFLOPs, but it comes at a modest accuracy of 82.8%. This suggests that while Early Exit 1 side-branch classifier is computationally efficient, it sacrifices some accuracy due to the lack of semantic information in earlier stages. In contrast, Early Exit 2 in EE-HiFuse achieves the highest accuracy in comparison with other configurations at 88.1%, with a moderate average inference time of 0.0629 seconds and 6.938 million parameters, making it an efficient compromise between performance and accuracy.

The EE-HiFuse model employing Early Exit 3 branch has an accuracy of 87.1%, with a balanced average inference time with 0.1269 seconds and number of parameters with 32.798 million.

In the final configuration of the EE-HiFuse model attains the same accuracy as Early Exit 2 with 88.1% but requires the highest average inference time with 0.1920 seconds and 18.144 GFLOPs, mirroring the number of parameters of the baseline. This configuration is the most computationally intensive, suggesting it is better suited for complex samples.

5. Conclusions

In this study, we introduced an early exit strategy for polyp colonoscopy image classification, integrating Convolutional Block Attention Modules (CBAM) into a baseline model. Our experiments on the Kvasir dataset validated the effectiveness of the proposed method, achieving superior classification accuracy while substantially reducing computational costs. These results highlight the potential of early exit strategies combined with attention mechanisms to enhance both performance and efficiency in medical image analysis. This work serves as a foundation for further exploration of efficient inference techniques in the field of medical imaging.

As a perspective, exploring the integration of more efficient side-branch classifiers with stronger learning capabilities. Furthermore, establishing a stopping schedule for early exit strategy is one of the most prevalent approaches.

Acknowledgments

We would like to sincerely thank the Faculty of Information and Communication Technologies, Hong Duc University provided professional support and created facilities to help us complete this research.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] B. Rokh, A. Azarpeyvand, and A. Khanteymoori, "A comprehensive survey on model quantization for deep neural networks in image classification," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 6, p. 50, 2023.
[2] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*, Chapman and Hall/CRC, 2022, pp. 291-326.
[3] A. Alqahtani, X. Xie, M. W. Jones, and E. Essa, "Pruning CNN filters via quantifying the importance of deep visual representations," *Comput. Vis. Image Underst.*, vol. 208, p. 103220, 2021.
[4] Y. Zhang *et al.*, "Advancing model pruning via bi-level optimization," in *Advances in Neural Information Processing Systems*, 2022.
[5] C. H. Wang, K. Y. Huang, Y. Yao, J. C. Chen, H. H. Shuai, and W. H. Cheng, "Lightweight deep learning: An overview," *IEEE Consum. Electron. Mag.*, 2022.
[6] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.
[7] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, 2016.
[8] Y. Matsubara, M. Levorato, and F. Restuccia, "Split computing and early exiting for deep learning applications: Survey and research challenges," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1-30, 2022.
[9] N. Passalis, J. Raitoharju, A. Tefas, and M. Gabbouj, "Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits," *Pattern Recognit.*, vol. 105, p. 107346, 2020.
[10] H. Li, H. Zhang, X. Qi, R. Yang, and G. Huang, "Improved techniques for training adaptive deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019.
[11] M. Phuong and C. H. Lampert, "Distillation-based training for multi-exit architectures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019.
[12] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama, "Adaptive neural networks for efficient inference," in *Proc. Int. Conf. Mach. Learn.*, 2017.
[13] J. Shen *et al.*, "Fractional skipping: Towards finer-grained dynamic CNN inference," in *Proc. AAAI Conf. Artif. Intell.*, 2020.
[14] G. Huang *et al.*, "Multi-scale dense networks for resource efficient image classification," *arXiv preprint arXiv:1703.09844*, 2017.
[15] T. K. Hu, T. Chen, H. Wang, and Z. Wang, "Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference," *arXiv preprint arXiv:2001.03460*, 2020.
[16] S. H. Gao *et al.*, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
[17] H. Zhang *et al.*, "ResNest: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.
[18] C. Srinivas *et al.*, "Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images," *J. Healthc. Eng.*, vol. 2022, p. 3264367, 2022.
[19] Z. Liu *et al.*, "Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network," *Knowl.-Based Syst.*, vol. 238, p. 107942, 2022.

[20] J. Cheng *et al*., "ResGANet: Residual group attention network for medical image classification and segmentation," *Med. Image Anal.*, vol. 76, p. 102313, 2022.
[21] A. Vaswani *et al*., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017.
[22] A. Dosovitskiy *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*.
[23] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2021.
[24] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.
[25] X. Huo *et al.*, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomed. Signal Process. Control*, vol. 87, p. 105534, 2024.
[26] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.
[27] K. Pogorelov *et al.*, "KVASIR: A multi-class image dataset for computer-aided gastrointestinal disease detection," in *Proc. 8th ACM Multimed. Syst. Conf.*, 2017.

Hoang Long Nguyen graduated at Hong Duc University in 2022, and received a Master's degree at Hong Duc University in 2024. His research interests include deep learning models, computer vision, medical image processing, and computational efficiency.
Email: nguyenhoanglong@hdu.edu.vn. ORCID: https://orcid.org/0009-0003-2327-4178

Minh-Vu Phan graduated at Hong Duc University in 2021, and received a Master's degree at Hong Duc University in 2024. His research interests include deep learning models, image processing.
Email: phanminhvu1997@gmail.com. ORCID: https://orcid.org/0009-0002-9872-5554

The-Anh Pham has been working at Hong Duc University as a permanent researcher since 2004. He received his PhD Thesis in 2013 from Francois Rabelais university in France. Starting from June 2014 to November 2015, he has worked as a full research fellow position at Polytech's Tours, France. He has then returned to Hong Duc University since 2016 and received the title of associate professor in 2019. His research interests include document image analysis, image compression, feature extraction and indexing, shape analysis and representation, and deep learning networks.
Email: phamtheanh@hdu.edu.vn. ORCID: https://orcid.org/0000-0002-0674-8066