Biomedical Signal Processing and Control 87 (2024) 105534

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

HiFuse: Hierarchical multi-scale feature fusion network for medical image classification

Xiangzuo Huo [a,b,1], Gang Sun [c,d,1], Shengwei Tian [a,*], Yan Wang [c,d,*], Long Yu [a], Jun Long [e], Wendong Zhang [a], Aolun Li [a,b]

[a] School of Information Science and Engineering, Xinjiang University, Urumqi, 830000, Xinjiang, China
[b] Xinjiang Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi, 830000, Xinjiang, China
[c] Department of Breast and Thyroid Surgery, The Affiliated Tumour Hospital of Xinjiang Medical University, Urumqi, 830011, Xinjiang, China
[d] Xinjiang Cancer Center/Key Laboratory of Oncology, The Affiliated Tumour Hospital of Xinjiang Medical University, Urumqi, 830011, Xinjiang, China
[e] Big Data Institute, Central South University, Changsha, 410083, Hunan, China

ARTICLE INFO

Keywords:
Medical image classification
Feature fusion
Swin-Transformer
Hybrid network
Multi-scale feature

ABSTRACT

Effective fusion of global and local multi-scale features is crucial for medical image classification. Medical images have many noisy, scattered features, intra-class variations, and inter-class similarities. Many studies have shown that global and local features are helpful to reduce noise interference in medical images. It is difficult to capture the global features of images due to the fixed size of the receptive domain of the convolution kernel. Although the self-attention-based Transformer can model long-range dependencies, it has high computational complexity and lacks local inductive bias. In this paper, we propose a three-branch hierarchical multi-scale feature fusion network structure termed as HiFuse, which can fuse multi-scale global and local features without destroying the respective modeling, thus improving the classification accuracy of various medical images. There are two key characteristics: (i) a parallel hierarchical structure consisting of global and local feature blocks; (ii) an adaptive hierarchical feature fusion block (HFF block) and inverted residual multi-layer perceptron(IRMLP). The advantage of this network structure lies in that the resulting representation is semantically richer and the local features and global representations can be effectively extracted at different semantic scales. Our proposed model's ACC and F1 values reached 85.85% and 75.32% on the ISIC2018 dataset, 86.12% and 86.13% on the Kvasir dataset, 76.88% and 76.31% on the Covid-19 dataset, 92.31% and 88.81% on the esophageal cancer pathology dataset. The HiFuse model performs the best compared to other advanced models. Our code is open source and available from https://github.com/huoxiangzuo/HiFuse.

1. Introduction

Medical image classification is a crucial task in computer-aided diagnosis, medical image retrieval, and mining. In recent years, Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in numerous medical image classification tasks [1–7]. However, due to the high similarity and detailed diversity in imaging modalities and clinicopathology, medical images exhibit significant intra-class variation and inter-class similarity, which necessitates the incorporation of global semantic information to improve performance.

Transformer [8] was originally used for modeling sequence to sequence prediction in natural language processing (NLP) tasks. Now Transformer has also attracted much attention in the computer vision community, ViT [9] by segmenting each image into patches with positional embeddings. Sequences of tokens are constructed, and a cascaded transformer block is applied to extract parameterized vectors as visual representations that model global semantic information through complex spatial transformations and long-range feature dependencies. Due to the lack of local spatial feature details. Li et al. proposed to utilize CNN feature maps as input tokens to capture feature neighborhood information [10]. However, they model the image as a one-dimensional sequence of tokens, ignoring the local inductive bias of the image, which affects the convergence speed and performance of the model.

Multi-scale global and local feature fusion is needed in image classification, segmentation, etc. Recent studies, such as ViTAE [11], StoHisNet [12], Transfuse [13], CMT [14], Comformer [15] and so on, which

* Corresponding authors.
E-mail addresses: huoxiangzuo@163.com (X. Huo), sung853219@163.com (G. Sun), tianshengwei@163.com (S. Tian), xjwangyan2012@163.com (Y. Wang).
[1] Equal contribution.

https://doi.org/10.1016/j.bspc.2023.105534
Received 24 February 2023; Received in revised form 15 August 2023; Accepted 24 September 2023
Available online 30 September 2023

Table 1
Acronyms involved in the text.

List of Acronyms	
CNN	Convolutional Neural Network
CT	Computed Tomography
DCNN	Deep Convolutional Neural Network
FFN	Feedforward Neural Network
HFF	Hierarchical Feature Fusion
IRMLP	Inverted Residual Multi-Layer Perceptron
LN	Layer Normalization
NLP	Natural Language Processing

solve the above problems to a certain extent by fusion of the features extracted by the convolution and self-attention mechanism [8]. By reviewing previous successful works, we believe that a successful classification model should have the following characteristics: (i) a robust backbone network, (ii) multi-scale global and local feature fusion, (iii) attention mechanism, (iv) relatively low computational complexity.

Inspired by ResNet [16], ConvNext [17], and Swin-Transformer [18], we propose a three-branch parallel hierarchical fusion network structure called HiFuse. we design new global and local feature blocks to extract global and local features in parallel respectively that fuse global and local representations at various semantic scales via HFF blocks, which makes our HiFuse more effective than previous classification methods that rely heavily on the Transformer. Experiments show that our method possesses better results on partial medical image datasets; in particular, our HiFuse-T outperforms Conformer-base-p16 in the case of a similar number of parameters using only about 1/3 (8.13 GFLOPs vs. 22.89 GFLOPs) of the computational cost. The acronyms appearing in this article are shown in Table 1.

Our advantages can be summarized as follows:

1. Combination and optimization of the characteristics that a good classification model should possess, and the proposal of a novel custom network framework termed as HiFuse. Global and local feature blocks efficiently capture local spatial features and global semantic information representations at different scales, respectively.

2. The proposal of an adaptive hierarchical feature fusion block (HFF block) to effectively fuse semantic information between different scale features of each branch. This block includes a spatial attention mechanism, a channel attention mechanism, an inverted residual multi-layer perceptron (IRMLP), and a shortcut connection.

3. The proposed HiFuse model achieves relatively good results on ISIC2018, Kvasir Covid-19-CT, and esophageal cancer pathology dataset.

2. Related work

Traditional medical image classification methods employ color, texture, shape, and combined descriptors. Baloch et al. [19] introduced a flexible skew-symmetric shape model to learn to identify latent variations within a defined neighborhood. Song et al. [20] presented a new texture descriptor that captures texture features by combining multi-scale Gabor transform and local binarized histograms for classifying lung tissue. Koitka et al. [1] manually extracted visual descriptors, which were then utilized for medical image classification.

Medical image classification based on deep learning has emerged in recent years. The deep convolutional neural network (DCNN) method that has greatly improved the classification accuracy and reduced the waste of resources for manual feature extraction is gradually applied to clinical auxiliary diagnosis [21,22]. Xu et al. [2] used DCNN to extract features from histopathological images of colon cancer and achieved good classification results. Koitka et al. [1] showed that transfer learning can achieve good results in medical image classification by fine-tuning the output of the last fully connected layer in a pre-trained ResNet-152 model. Kumar et al. [6] proposed an approach of integrating two pre-trained DCNN architectures to create a more powerful classifier. Shen et al. [3] introduced a multi-scale crop pooling strategy for DCNN to capture lung nodule classification features in chest CT images. Esteva et al. [4] developed a model that is trained on 129,450 clinical images to diagnose the most common and deadly skin cancers, which achieved performance comparable to that of 21 dermatologists. Cheng et al. [23] proposed a modular group attention block that captures feature correlations in medical images' channel and space dimensions. However, these models cannot collect sufficient contextual information, and global semantic information features are equally important in high-resolution medical images.

Transformer is used for vision. Transformer [8] was originally used in NLP. It extracts intrinsic properties through a self-attention method. ViT [9], as a pioneer work, verifies the feasibility of pure Transformer architecture in computer vision tasks [24–26], and is gradually used for image classification [10,27,28], object detection [29–31], semantic segmentation [13,32,33], image enhancement [34] and image generation [35]. Researchers [36–39] have tried various approaches to make transformers more successful in computer vision. However, the self-attention mechanism in the visual Transformer has been found to often ignore local feature details. This can lead to difficulties in distinguishing the object of interest from the background, particularly when the local features are weak.

Multi-scale global and local feature fusion. To address the lack of local features, DeiT [28] proposed using distilled tokens to transfer the CNN-based features to the visual Transformer. T2TViT [10] proposed using a tokenization module recursively reorganizing images to consider adjacent pixels' tokens. The models, such as VitAE, StoHis, CMT, Conformer, Focal and so on, [11,12,14,15] not only inherit the structural advantages of CNN and Transformer, but also verify that the coupling of local features and global representation can significantly enhance Transformer discriminability of weak local features. The above models perform well on natural datasets such as ImageNet and various downstream tasks, however, when applied to the medical image domain, the results are unsatisfactory. Because the datasets of medical images are insufficient, pathological features are more scattered and difficult to discover than those of ordinary images.

In the field of medical analysis, TransFuse [13], GasHis Transformer [40] and Medical Transformer [36] combined Transformer and CNN for fusion, achieved good results in downstream tasks such as segmentation or detection, and proved that capturing global and local features at the same time can reduce noise interference on medical images to a certain extent. Different from the above studies, we intend to design a general-purpose backbone for robust medical feature extraction, rather than only for downstream tasks of a certain class of medical images such as segmentation or detection. The salient point of our proposed HiFuse is to efficiently extract and fuse global and local features on the encoder, parallel branches can maintain their respective modeling without introducing new noise. The fusion of different hierarchical of local and global in HIGF-Net [41] can capture high-level semantic information and low-level spatial features. HFF blocks subtly interact hierarchically and suppress the introduction of noise information, promoting the fusion of global and local features. We believe this is the key to solving general medical image analysis tasks without over-improving the decoder or segmentation head or detection head.

Therefore, we decide to take full advantage of multi-scale global–local features and a robust fusion network to fuse them. Our proposed HiFuse model defines a three-branch hierarchical parallel fusion structure, parallel design to keep global and local branches from interfering with each other, and designs a hierarchical feature fusion block (HFF block) to fuse these features and keep global and local branches undisturbed. HiFuse inherits not only the advantages of CNN and Transformer but also the local features and global representations coupled at different scales.

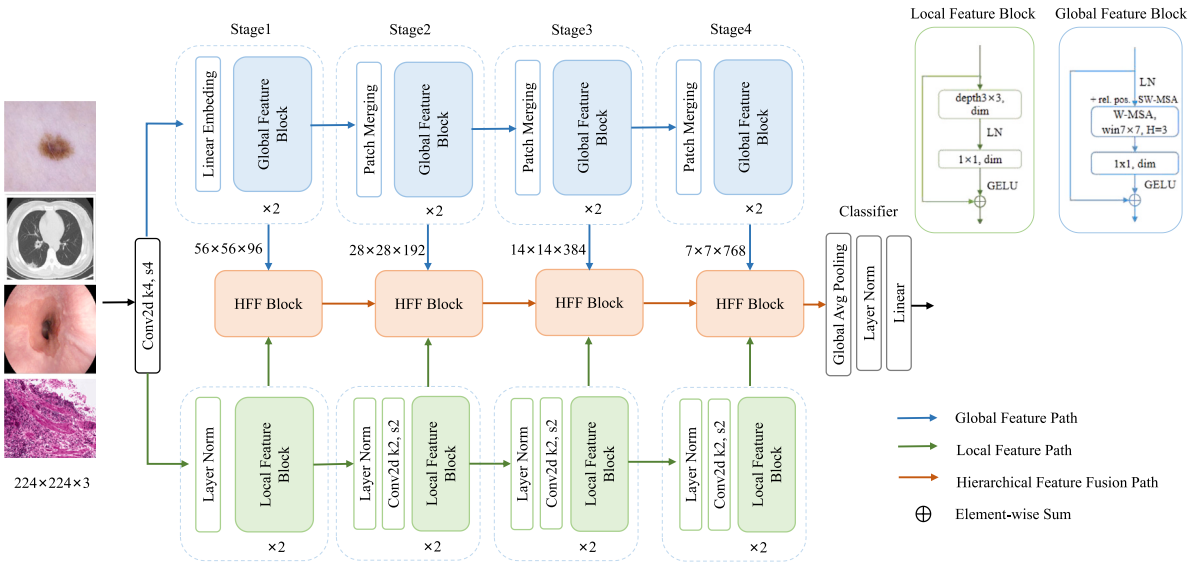

Fig. 1. Overall structure of the HiFuse model.

3. Proposed method

The HiFuse model, as a new medical image classification method, is proposed to effectively obtain local spatial information and global semantic representations of medical images at different scales. We use a parallel structure to extract the global and local information of medical images from the global and local feature blocks, fusing the features of different hierarchies through the HFF block, downsampling step, and finally, obtaining the classification result. In the following sections, we first introduce the overall structure of the HiFuse model, then introduce the global feature block and the local feature block, respectively. We describe the details of the HFF block in Section 3.5.

3.1. Stem

Image into a stem, convolved by a stride size of 4 and a kernel size of 4 to divide the image into 4 × 4 Patches, each with H × W of 56 × 56 (i.e., the resolution is reduced to 1/4), and then input the global path after linear embedding and the local path after LayerNorm [42], respectively.

3.2. The multi-stage design for HiFuse

In order to improve the accuracy of the classification model of medical images, it is necessary to fuse local features and global representations from different hierarchical levels. We designed a parallel network structure for hierarchical feature fusion. The overall structure of HiFuse is shown in Fig. 1. The local branch is used to extract the local features of the image, and the global branch is used to extract the global semantic representation of the image. Global branches are downsampled by patch merging [18] and then input global feature blocks for feature transformation. Local branches are downsampled by convolution with a stride size of 2 and a convolution kernel of 2. The specific parameters are shown in Table 2.

The three-branch parallel structure means that local features and global representations can be preserved to the greatest extent without interfering with each other. Feature maps of different hierarchical levels are constructed through four stages. We believe that a multi-branch structure enhances the feature representation of the model, and the hierarchical structure is robust to the multi-scale features of the model and downstream task friendly. HFF blocks are used to fuse each stage's local features and global representations and connect the output from the previous stage. Finally, the combined features are fed to the linear classifiers of global average pooling and LayerNorm for classification. We build different HiFuse variants, HiFuse-Tiny/Small/Base; these variants have different numbers of global and local feature blocks in each stage and build models with different depths to deal with datasets of various sizes. The hyper-parameters of these model variants are:

- HiFuse-Tiny: Block numbers = (2, 2, 2, 2)
- HiFuse-Small: Block numbers = (2, 2, 6, 2)
- HiFuse-Base: Block numbers = (2, 2, 18, 2)

3.3. Global feature block

The imaging methods and clinical pathology of medical images are diverse, with significant intra-class changes and inter-class similarities. The acquisition of global semantic information is very important. Therefore, we introduced the Windows Multi-head Self-Attention (W-MSA) in the global feature extraction branch. W-MSA is the Swin-Transformer [18] first proposed, Compared with the Multi-head Self-Attention (MSA) module in the Transformer, the W-MSA module, which can effectively reduce the amount of computation and divide the feature map into $M \times M$ sizes. Window one by one, and then perform self-attention on each Window individually. The computational complexity formula is shown in (1).

$$\begin{aligned} \Omega(\text{MSA}) &= 4hwC^2 + 2(hw)^2C \\ \Omega(\text{W-MSA}) &= 4hwC^2 + 2M^2hwC \end{aligned} \tag{1}$$

where h represents the height of the feature map, w represents the width of the feature map, C represents the depth of the feature map, and M represents the size of each window.

For each stage, by incorporating the patch into the global feature block, the feature map goes through LayerNorm layer into W-MSA and then through the linear layer with the GELU activation function, as shown in Fig. 1. The common self-attention mechanism post-connected to the FFN aims to introduce nonlinearity, and nonlinear activation is usually performed on the expanded channel dimension generated by the linear layer to increase the representative capability of the model. To further reduce the computational cost, we replace the FFN with a GELU activation function doing only the nonlinear transformation while compensating in the HFF block. A residual connection is applied after each module, a relative position bias (rel. pos.) is used, and Shift W-MSA is introduced in the next module. This process is depicted in (2).

$$\begin{aligned} g_i &= f^{1\times1}\left(\text{W-MSA}\left(LN\left(G_{i-1}\right)\right)\right) + G_{i-1} \\ G_i &= f^{1\times1}\left(\text{SW-MSA}\left(LN\left(g_i\right)\right)\right) + g_i \end{aligned} \tag{2}$$

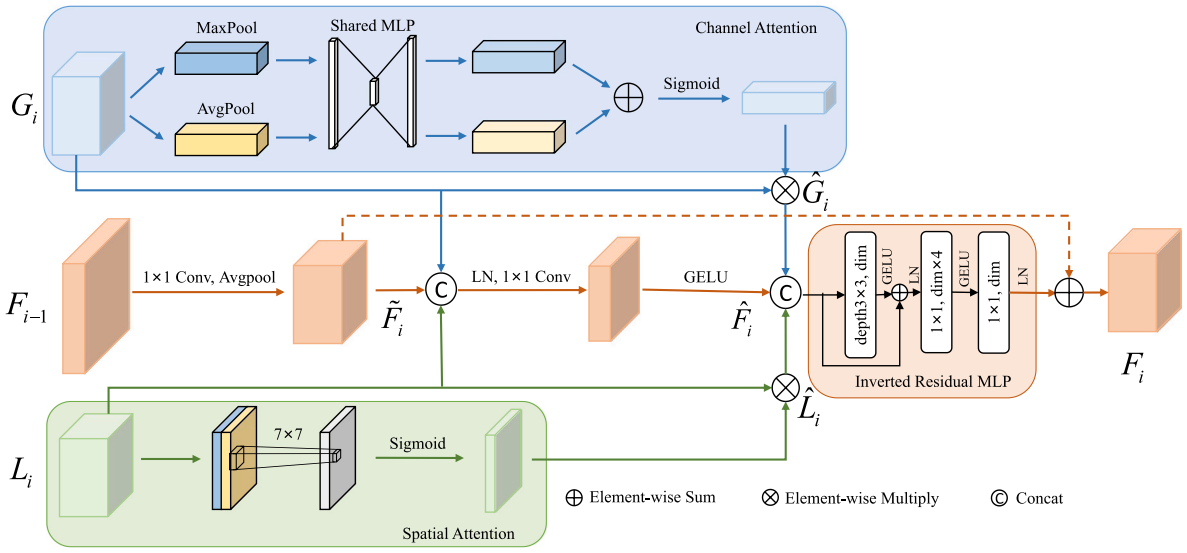

Fig. 2. HFF block detail display.

where G_i and g_i denote the output features of the Shift W-MSA and the W-MSA for the global feature block. $f^{1\times1}$ is the convolution operation with a convolution kernel size of 1×1, which is equivalent to the linear operation. LN is the LayerNorm operation. Finally, the extracted global features are input into the HFF block.

3.4. Local feature block

Local spatial features in medical images are also very important. The local feature block, shown in Fig. 1, uses a 3×3 depthwise convolution [43,44], a special case of grouped convolutions [45]; the number of groupings is equal to the number of channels. The use of depthwise convolution effectively reduces the FLOPs of the network. And then, cross-channel information interaction through linear layers. Finally, the extracted local features are input into the HFF block. This process is depicted in (3).

$$L_i = f^{1\times1}\left(LN\left(f^{depth3\times3}(L_{i-1})\right)\right) + L_{i-1} \tag{3}$$

where L_i denotes the output features of the local feature block. $f^{depth3\times3}$ is the depthwise convolution operation with a convolution kernel size of 3×3.

Macroscopically, global and local branch structures are similar, and the design of the same number of channels and hierarchical structure lays the foundation for fusing global and local encoding features of different scales. How to effectively adapt and fuse features of different scales in each branch becomes a new problem [46]. To this end, we propose the HFF block.

3.5. HFF block

Adaptive hierarchical feature fusion block can adaptively fuse local features from different layers, global representations, and semantic information after fusion from the previous hierarchy according to the input features, as shown in Fig. 2. Among them, G_i denotes the matrix of features produced by the global feature block, L_i represents the feature matrix output by the local feature block, F_{i-1} denotes the matrix of features generated by the previous stage of HFF, and F_i represents the feature matrix generated by HFF fusion at this stage. For the structure comparison of the structure of ResNet, Swin-Transformer, ConvneXt, and our HFF blocks, as shown in Fig. 3. The feature fusion operation uses the following formula:

$$\begin{aligned}
\hat{G}_i &= \mathrm{CA}\left(G_i\right) \otimes G_i \\
\hat{L}_i &= \mathrm{SA}\left(L_i\right) \otimes L_i \\
\tilde{F}_i &= \mathrm{Avgpool}\left(f^{1\times1}\left(F_{i-1}\right)\right) \\
\hat{F}_i &= f^{1\times1}\left(\mathrm{Concat}\left[G_i, L_i, \tilde{F}_i\right]\right) \\
F_i &= \mathrm{IRMLP}\left(LN\left(\mathrm{Concat}\left[\hat{G}_i, \hat{L}_i, \hat{F}_i\right]\right)\right) + \tilde{F}_i
\end{aligned} \tag{4}$$

where $\otimes$ represents element-wise multiply, $\hat{G}_i$ is generated by the combination of channel attention, $\hat{L}_i$ is generated by the combination of spatial attention, and $\tilde{F}_i$ is generated downsampled by the previous stage of the HFF block. $\hat{F}_i$ is the result of global–local features and the fusion of the previous stage. Finally, $\hat{F}_i$, $\hat{G}_i$ and $\hat{L}_i$ are concatenated and generate feature F_i through a IRMLP.

3.5.1. Attention mechanisms

The attention mechanism is an adaptive selection process that selectively focuses on relevant regions by assigning importance to different parts of the input. The channel attention mechanism, in particular, allows the network to selectively focus on important features and objects, which has been shown to be crucial in previous research [47,48], and spatial attention focuses on important spatial regions [49–51].

Self-attention in global feature blocks can capture global information in space and time to some extent [52]. All tokens are propagated forward point-wise. However, recently popular visual Transformer [9, 18,53–55] usually ignore the adaptation of channel dimensions. Therefore, the HFF block feeds the global features into the channel attention (CA) mechanism, which exploits the interdependencies between channel maps to improve the feature representation for specific semantics. The local features are input into the spatial attention (SA) mechanism in order to selectively enhance important regions and suppress irrelevant information, thus allowing for improved preservation of important local details.

$$\begin{aligned}
\mathrm{CA}(x) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(x)) + \mathrm{MLP}(\mathrm{MaxPool}(x))) \\
\mathrm{SA}(x) &= \sigma\left(f^{7\times7}(\mathrm{Concat}[\mathrm{AvgPool}(x), \mathrm{MaxPool}(x)])\right)
\end{aligned} \tag{5}$$

where σ is the Sigmoid function, $f^{7\times7}$ is the convolution operation with a convolution kernel size of 7×7.

The output of local features at each level via spatial attention is combined with the output of global features at each level via channel attention and finally fed to the IRMLP module.

Table 2
HiFuse specific parameters.

Stage	Output size	Local branch		HFF branch	Global branch	
stem	56 × 56, 96	4 × 4, 96, stride 4		–	4 × 4, 96, stride 4	
1	56 × 56	depth 3 × 3, 96 1 × 1, 96	× 2	→ spatial attention channel attention← depth 3 × 3, 96 1 × 1, 384 1 × 1, 96	MSA, win7 × 7, head 3, rel. pos. SW-MSA 1 × 1, 96	× 2
2	28 × 28	depth 3 × 3, 192 1 × 1, 192	× 2	→ spatial attention channel attention← depth 3 × 3, 192 1 × 1, 768 1 × 1, 192	MSA, win7 × 7, head 6, rel. pos. SW-MSA 1 × 1, 192	× 2
3	14 × 14	depth 3 × 3, 384 1 × 1, 384	× 2	→ spatial attention channel attention← depth 3 × 3, 384 1 × 1, 1536 1 × 1, 384	MSA, win7 × 7, head 12, rel. pos. SW-MSA 1 × 1, 384	× 2
4	7 × 7	depth 3 × 3, 768 1 × 1, 768	× 2	→ spatial attention channel attention← depth 3 × 3, 768 1 × 1, 3072 1 × 1, 768	MSA, win7 × 7, head 24, rel. pos. SW-MSA 1 × 1, 768	× 2
Parameters				**82.49 M**		
FLOPs				**8.13 G**		

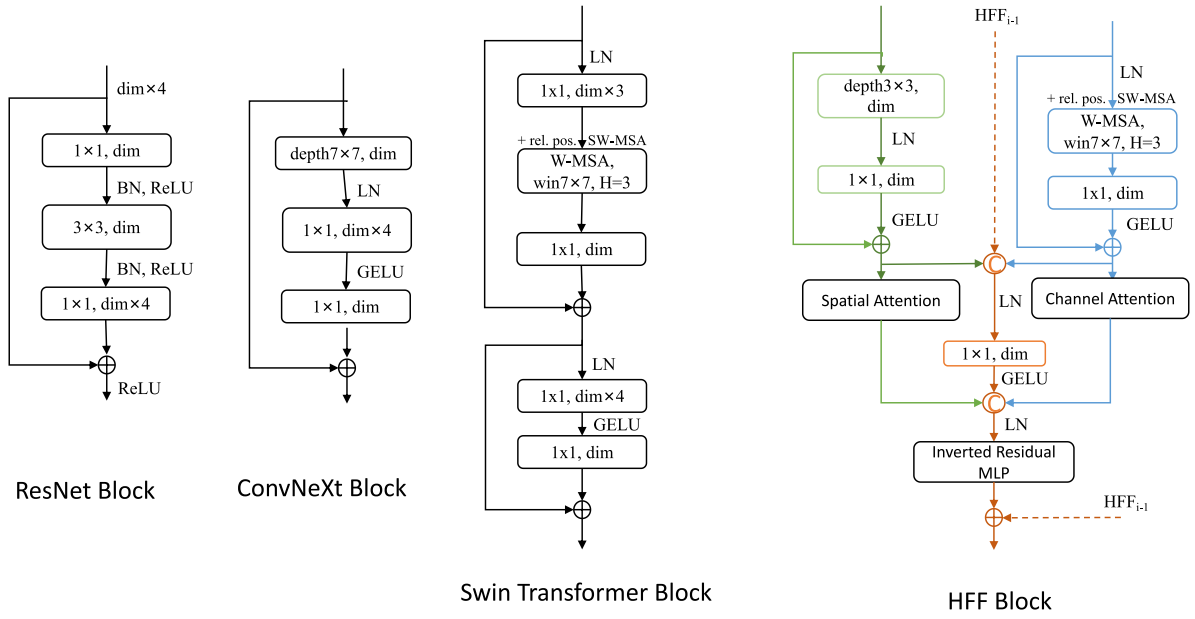

Fig. 3. Block designs for a ResNet, a ConvNeXt, a Swin-Transformer and a HFF.

3.5.2. IRMLP

The design of the inverted residual MLP (IRMLP) is shown in Fig. 2, consists of a 3 × 3 depthwise convolution with residuals and 2 linear transformation layers with nonlinear transformations by GELU in the expanded channel dimension, which compensates for the effects caused by the missing FFN and, to some extent, prevents gradient disappearance, explosion and network degradation, thus effectively capturing the global and local feature information at each level. This process is described in (6).

$$\mathrm{IRMLP}(x) = f^{1\times1}\left(f^{1\times1}(f^{depth3\times3}(x) + x)\right) \tag{6}$$

4. Experimental result

In order to assess the effectiveness and reliability of the HiFuse model, we conducted a series of experiments utilizing four datasets. The results of these experiments demonstrate that our approach outperforms previous state-of-the-art network models in terms of classification performance. In the following sections, we will provide a comprehensive overview of the datasets used, the evaluation metrics employed, and the experimental setup. Subsequently, we will present the results of our experiments on all datasets, including a thorough analysis of the performance of our model through a series of ablation studies on the ISIC2018 dataset.

4.1. Dataset

ISIC2018 [56]: We utilize the ISIC2018 dataset for our experiments, which is specifically designed for the task of skin lesion diagnosis. The dataset comprises of 10,015 images from seven different categories, including melanocytic nevi, dermatofibroma, melanoma, actinic keratosis, benign keratosis, basal cell carcinoma, and vascular lesions. The original image size in the dataset is 650 × 450 pixels, which we downscaled to 224 × 224 pixels for our experiments. We followed the data division method proposed by Chen [23] and used 70% of

the samples (7010) for training and validation, while reserving the remaining 30% of the samples (3005) for testing.

Kvasir [57]: The dataset comprises of 4000 endoscopic gastrointestinal diseases and is divided into eight classes, each comprising of 500 images. The dataset includes various images in each category, featuring anatomical landmarks (such as Z-line, pylorus, or cecum) and pathological findings (such as esophagitis, polyps, or ulcerative colitis). The dataset comprises of images with different resolutions ranging from 720×576 to 1920×1072 pixels. We downscale all images to 224×224 pixels, follow the data division method in literature [57], and split the dataset into a 50:50 ratio with 2-fold cross-validation.

COVID-19-CT [58]: The COVID19-CT dataset used in this study includes 746 CT scan images, with 349 being positive for COVID-19 and 397 being negative or showing other types of diseases. The image sizes in the dataset range from 143×76 to 1637×1225 pixels. To standardize the images, they were all scaled to 224 x 224 pixels. The dataset was divided into training, validation, and testing sets following the method outlined in literature [58], with a ratio of 0.6:0.15:0.25 respectively.

Esophageal cancer pathology dataset: The dataset was sourced from 50 subjects that were admitted to the Tumor Hospital of Xinjiang Medical University and contained stage 1 A, stage 1B, stage 2 A, and stage 2B. The dataset was a crop with a 1024×896 pixels resolution, and all images were downscaled to 224×224 pixels. With the help of two medical imaging experts, we obtained 524 slices color-normalized by the Vahadane [59] method. The experiment was validated using a 5-fold crossover.

4.2. Evaluation metrics and implementation details

We choose ACC, F1, Precision, Recall, MCC, Kappa, and AUC as classification indicators. These metrics are calculated based on the confusion matrix. The symbols used in the confusion matrix are defined as follows: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Therefore, Accuracy (ACC) is calculated by Eq. (7) to get the percentage of correctly identified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Use (8) to calculate the precision rate, the proportion of samples with correct, true values among the samples predicted to be correct, to reflect the accuracy of the model prediction.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Use (9) to calculate the recall rate, the number of positive samples are found in the data for which all true values are predicted correctly, to reflect the comprehensiveness of the model prediction.

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

The definition of the F1 score formula for each category is shown in (10). F1 score can solve the balance between precision and Recall, and the higher the value, the better.

Matthews Correlation Coefficient is a metric used to evaluate binary classification problems with a value between −1 and 1, where 1 indicates a completely correct prediction, 0 indicates a random prediction, and −1 indicates a completely incorrect prediction. The formula for calculating MCC is as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

Kappa coefficient is also a metric used to evaluate the performance of classification models, and it is mainly used to measure the consistency of classifiers. The value of Kappa coefficient ranges from 0 to 1, where 1 indicates perfect agreement and 0 indicates random agreement. The formula for calculating Kappa coefficient is as follows:

$$kappa = \frac{(Po - Pe)}{(1 - Pe)} \tag{12}$$

where Po denotes the observed consistency, that is, the actual accuracy, and Pe denotes the random consistency, that is, the accuracy of random guesses.

We implement our PyTorch-based approach by training on an NVIDIA RTX 3090 GPU with 24 GB of video memory. The base learning rate value is 1e-4, the batch size is 32, and the cosine annealing learning rate strategy is adopted. To ensure the fairness of the experiments, we adopt an image size of 224×224, maintain consistency in the operating environment and hyperparameters, and utilize the same training, validation, and test sets as outlined in previous literature. We were conducting experiments under the mmcv [60] framework. We employ the softmax function as the output layer and utilize the categorical cross-entropy loss function to compute the loss value:

$$CrossEntropyLoss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{k} y_i^t \log y_i^p \tag{13}$$

where N represents the total number of samples, K represents the number of categories, y_i^t is the target label, and y_i^p is the model's predicted value output. More parameter settings are shown in Table 3.

Table 3
Experimental setting.

Training config	224 × 224
optimizer	AdamW
drop path rate	0
base learning rate	1e−4
min learning rate	1e−6
weight decay	0.01
optimizer momentum	$\beta 1$, $\beta 2$ = 0.9,0.999
batch size	32
training epochs	100
learning rate schedule	CosineAnnealing
warm up schedule	linear
warm up epochs	1

4.3. Comparison with other advanced models

To evaluate the performance of the HiFuse model, we conducted a comparison with state-of-the-art network models in recent years, including VGG [61], Mlp-Mixer [62], ViT [9], T2TViT [10], DeiT [67], Swin-Transformer [18], Conformer [15], ConvNeXt [17], PerViT [63], Focal [64], UniFormer [65], and BiFormer [66]. Experiments were conducted using four datasets with the same parameter settings, and the effectiveness of the proposed network was validated by comparing its test set results to other models.

4.3.1. Results on ISIC2018 dataset

Table 4 illustrates the evaluation of the proposed model and several state-of-the-art classification algorithms on the ISIC2018 dataset. The evaluation is conducted using the data partitioning method outlined in [23] and training from scratch.

As what can be seen from the table, HiFuse has significant advantages in medical image classification. Like ConvNeXt and Swin-Transformer, our approach utilizes a hierarchical structure to enhance the feature representation capabilities of neural networks across different scales. However, it is noteworthy that our model achieves 5.9% and 6.06% improvement in classification accuracy over ConvNeXt and Swin-Transformer respectively (79.95% vs 85.85% and 79.79% vs 85.85%). Like Conformer, a multi-branch structure combines the advantages of CNN and Transformer. However, the difference lies in that we do not interact the information of the branches but fuse the

Table 4
Performance comparison of the ISIC2018 dataset.

Method	Params(M)	FLOPs(G)	Acc%	F1%	Prec%	Recall%
VGG-19 (ICLR2015) [61]	143.68	19.67	79.25	61.83	63.71	60.89
Mixer-L/16 (NIPS2021) [62]	208.20	44.57	78.92	59.88	61.36	59.16
T2T-ViT_t-24 (NIPS2021) [10]	64.00	12.69	77.59	57.21	59.60	55.94
DeiT-base (ICML2021) [28]	86.57	16.86	72.31	41.01	47.19	44.09
VIT-B/16 (ICLR2021) [9]	86.86	33.03	78.32	60.93	64.16	60.52
VIT-B/32 (ICLR2021) [9]	88.30	8.56	77.92	57.52	58.74	56.90
Swin-B (ICCV2021) [18]	87.77	15.14	79.79	63.95	65.09	63.65
Conformer-base-p16 (ICCV2021) [15]	83.29	22.89	82.66	72.44	73.31	71.66
ConvNeXt-B (CVPR2022) [17]	88.59	15.36	79.95	63.24	64.90	62.06
PerViT-M (NIPS2022) [63]	43.04	9.00	81.64	67.66	68.19	67.29
Focal-B (NIPS2022) [64]	87.10	15.30	79.64	62.88	65,73	60.68
UniFormer-B (TPAMI2023) [65]	50.02	8.30	82.44	68.41	70.67	66.54
BiFormer-B (CVPR2023) [66]	56.04	9.80	82.66	68.95	72.66	66.47
HiFuse-Tiny	82.49	8.13	82.99	72.99	73.67	72.87
HiFuse-Small	93.82	8.84	83.59	72.70	72.70	73.14
HiFuse-Base	127.80	10.97	**85.85**	**75.32**	**74.57**	**76.58**

Table 5
Performance comparison of the Kvasir dataset.

Method	Acc%	F1%	Prec%	Recall%
VGG-19	77.75	77.75	77.86	77.83
Mixer-L/16	74.30	74.14	74.43	74.34
T2T-ViT_t-24	76.90	76.78	77.60	76.91
DeiT-base	52.15	48.48	56.72	52.29
VIT-B/16	76.10	75.94	76.49	76.23
VIT-B/32	73.80	73.50	74.24	73.72
Swin-B	77.30	77.29	77.74	77.44
Conformer-base-p16	84.25	84.27	84.45	84.37
ConvNeXt-B	74.62	74.41	75.69	74.62
PerViT-M	82,40	82.30	82.88	82.40
Focal-B	78.00	77.93	78.19	78.01
UniFormer	83.10	83.04	83.09	83.10
BiFormer	84.25	84.26	84.67	84.25
HiFuse-Tiny	84.85	84.89	84.96	84.90
HiFuse-Small	**86.12**	**86.13**	**86.25**	**86.13**
HiFuse-Base	85.97	86.07	86.29	86.01

Table 6
Performance comparison of the Covid19 dataset.

Method	Acc%	F1%	Prec%	Recall%
VGG-19	59.14	57.55	59.04	58.13
Mixer-L/16	70.43	70.12	70.38	70.06
T2T-ViT_t-24	63.44	60.34	65.68	61.89
DeiT-base	50.54	39.31	44.47	47.96
VIT-B/16	65.05	64.88	64.90	64.87
Swin-B	60.75	56.36	63.20	58.95
VIT-B/32	61.83	60.59	61.89	60.94
Conformer-base-p16	75.81	75.60	76.81	77.81
ConvNeXt-B	55.38	54.68	54.95	54.81
PerViT-M	74.75	73.95	75.96	73.93
Focal-B	69.36	68.06	70.66	68.35
UniFormer	71.38	71.35	72.18	71.89
BiFormer	72.05	71.95	71.94	71.96
HiFuse-Tiny	74.73	74.67	74.65	74.73
HiFuse-Small	**76.88**	**76.31**	**77.78**	**76.19**
HiFuse-Base	76.34	76.17	76.30	76.11

branches of the HFF block at different levels to reduce the computational complexity while improving the accuracy of medical image classification.

The experimental results presented in this study provide evidence for the effectiveness of the hierarchical fusion of feature information from multiple branches in reducing computational complexity while also improving the performance of the classification model. Our proposed method, HiFuse, demonstrates these characteristics in varying degrees. The FLOPs of HiFuse-Base is only 10.97 G, which has the best classification accuracy (85.85%) on the ISIC2018 dataset.

4.3.2. Results on Kvasir dataset

In order to further explore the generalization ability of HiFuse, we conduct experiments in the Kvasir dataset according to the division method and 2-fold cross-validation in the literature [57], and the final results are averaged (see Table 5). It can be seen that HiFuse-Small's accuracy (86.12%) and F1 value (86.13%) achieve the best performance in this dataset. Moreover, the accuracy of HiFuse-Base decreases slightly due to the increase in depth, which can be improved slightly by setting the appropriate drop path hyperparameter (the table ensures a fair comparison, not shown). The observation highlights the possibility that utilizing HiFuse-Small, which has a shallower depth, may result in improved classification performance when applied to other smaller medical image classification datasets.

4.3.3. Results on Covid-19-CT dataset

As shown in Table 6, we evaluate the COVID-19-CT dataset using the data partitioning method from the literature [39] and training network from scratch. Among them, the bold represents the best performance.

DeiT [28] has limited classification performance on this dataset. Our analysis of the situation suggests that the network structure, optimized for large datasets like ImageNet, may not be suitable for small datasets like COVID19-CT. It can be seen from the experimental results that the pure convolutional models ConvNeXt [17] and VGG [61] are not good at extracting features from small-sample CT datasets, while PerViT extracts coarse-grained global and fine-grained local features, HiFuse and Conformer using CNN and self-attention hybrid network structures have higher classification performance.

HiFuse-Small's accuracy (76.88%) and F1 value (76.31%) achieved the best performance on this dataset. The results indicate that HiFuse can be versatile and remain effective in other similar applications while maintaining a high level of classification performance.

4.3.4. Results on Esophageal Cancer Pathology dataset

In this dataset, the overall sample size is small, and the long-tail problem is obvious, so we set 300 epochs, other parameters remain the same, and the experimental results are the results after the five-fold cross-validation average (see Table 7). It can be seen that the HiFuse series perform the best, and many models perform unsatisfactorily on the clinical dataset and even show model non-convergence. Similar to the findings on the Kvasir dataset, the multiscale global–local feature fusion model performs better, compared with other models in a real

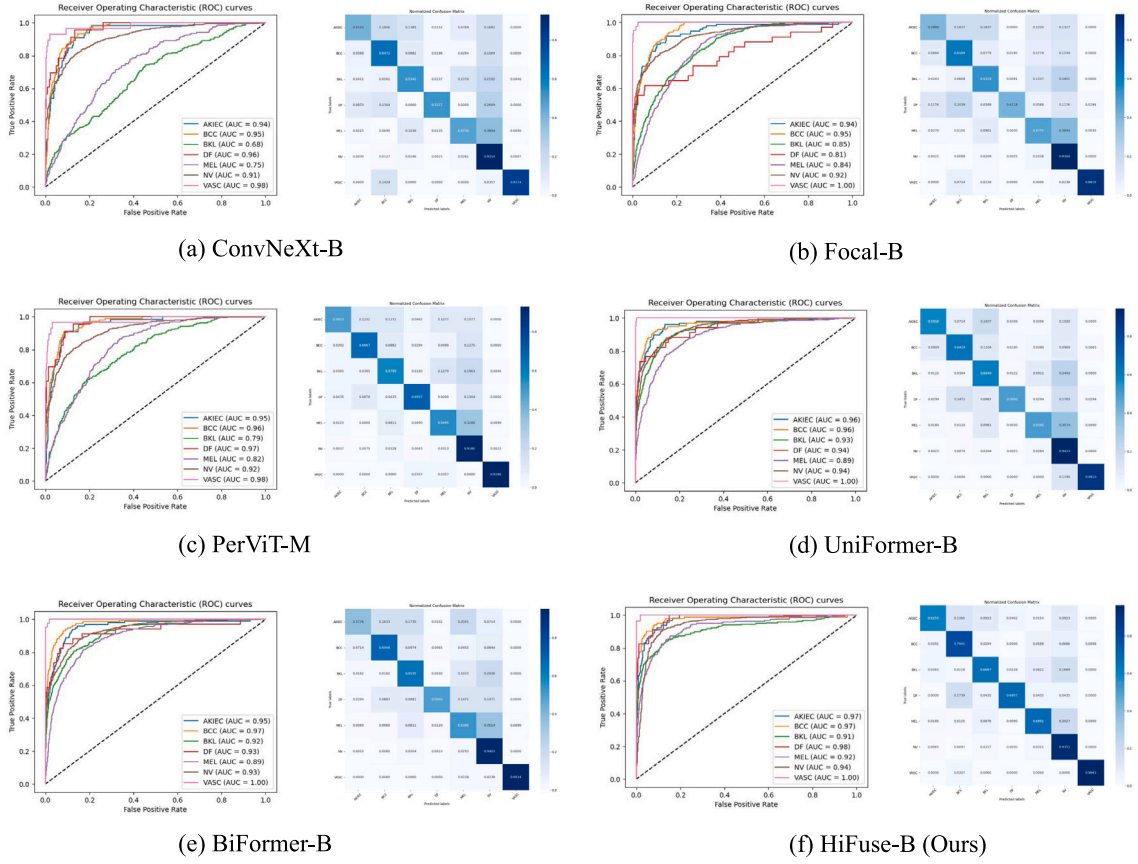

Fig. 4. ROC curve and confusion matrix on ISIC2018 dataset.

Table 7
Performance comparison of the esophageal cancer pathology dataset.

Method	Acc%	F1%	Prec%	Recall%
VGG-19	69.23	44.22	47.04	44.49
Mixer-L/16	77.88	60.44	64.58	60.33
T2T-ViT_t-24	65.38	13.79	14.50	15.28
DeiT-base	63.46	19.41	15.87	25.29
VIT-B/16	66.35	51.04	55.17	48.81
Swin-B	64.42	24.53	41.02	27.78
VIT-B/32	65.38	29.77	53.91	31.43
Conformer-base-p16	90.30	86.13	84.39	88.8
ConvNeXt-B	68.27	36.65	39.09	36.92
PerViT-M	70.59	52.66	58.43	50.33
Focal-B	80.47	80.31	80.52	80.22
UniFormer	81.37	70.96	72.38	69.97
BiFormer	78.43	66.87	68.07	66.40
HiFuse-Tiny	91.35	87.37	89.03	85.91
HiFuse-Small	**92.31**	**88.81**	**89.14**	**89.56**
HiFuse-Base	88.46	80.82	84.02	78.45

clinical classification task on small datasets, and demonstrates that our model can be robust on different medical datasets and can mitigate the impact of the long-tail problem to some extent.

4.3.5. ROC curves and confusion matrix

We further compared the ROC curves, confusion matrix, MCC, and kappa, and AUC of the 2022–2023 SOTA classification model on ISIC2018 and Kvasir datasets for a more comprehensive evaluation.

It can be seen from Figs. 4 and 5 that in the ISIC2018 data set, HiFuse can also classify the MEL and NV categories that are more likely to be misclassified, and the AKIEC category is more accurate than other models. In the Kvasir dataset, HiFuse has a stronger discrimination ability for the two categories of dyed-lifted-polyps and dyed-resection-margins, and the overall accuracy is higher.

Table 8
Recent model performance of MCC, Kappa and AUC indicators is compared in ISIC2018 and Kvasir datasets.

	ISIC2018			Kvasir		
	MCC	Kappa	AUC	MCC	Kappa	AUC
ConvNeXt-B	0.6022	0.6005	0.8831	0.7119	0.7100	0.9343
PerViT-M	0.6412	0.6406	0.9263	0.7998	0.7989	0.9756
Focal-B	0.5964	0.5949	0.9017	0.7490	0.7486	0.9606
UniFormer-B	0.6505	0.6482	0.9457	0.8070	0.8069	0.9778
BiFormer-B	0.6604	0.6592	0.9428	0.8205	0.8200	0.9712
HiFuse-T	0.6639	0.6619	0.9386	0.8055	0.8051	0.9729
HiFuse-S	0.6896	0.6892	0.9439	**0.8416**	**0.8414**	**0.9820**
HiFuse-B	**0.7282**	**0.7280**	**0.9585**	0.8402	0.8400	0.9815

As can be seen from the Table 8, compared with other recent methods, HiFuse shows excellent performance on MCC, Kappa and AUC indicators, and has better robustness.

4.4. Ablation study

As shown in the Table 9, we evaluated the impact of each component on the model on the ISIC2018 dataset, starting from the local path, adding the global path, channel & spatial attention, inverted residual MLP, and shortcut, finally, to form the final HiFuse-Tiny model, after adding the global block Acc and F1 increased by 2.47% and 10.2%. After adding components in HFFblock, Acc and F1 increased by 7.4% and 8.67%. It can also be seen that combining the global features can significantly improve the representation ability of the model, and the

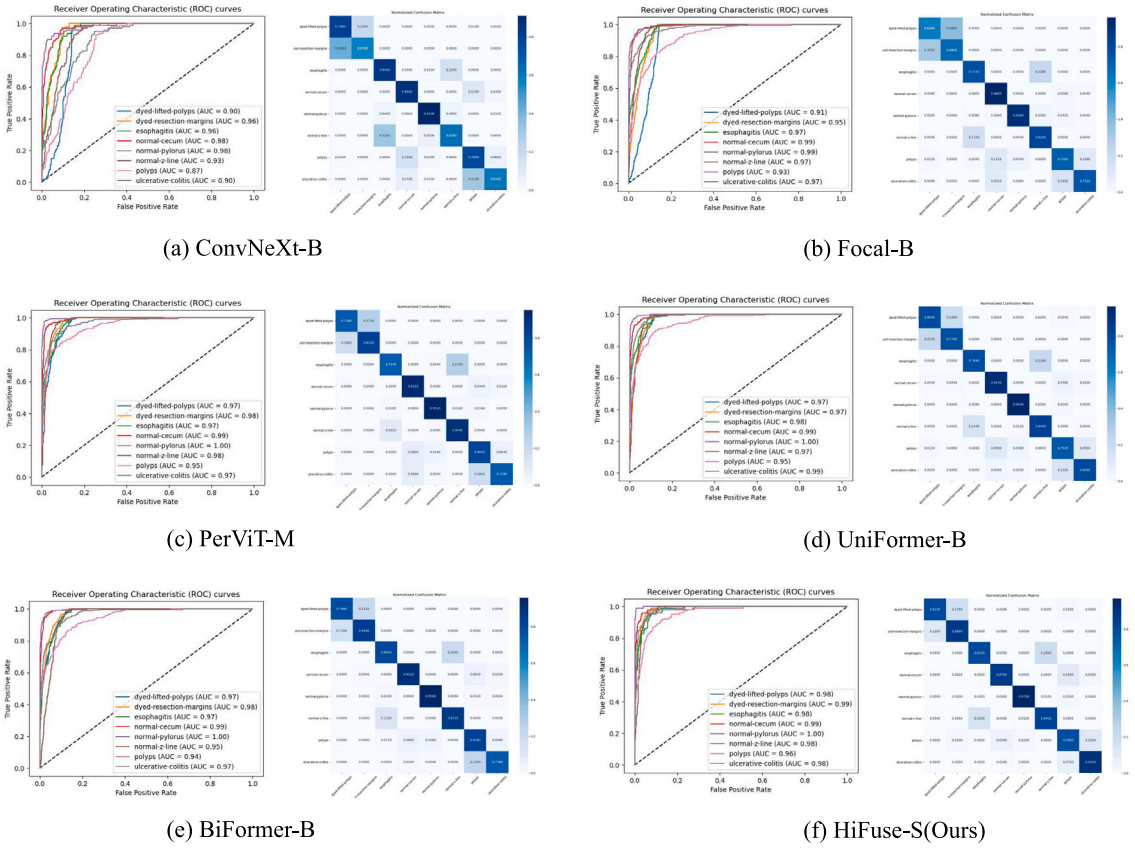

Fig. 5. ROC curve and confusion matrix on Kvasir dataset.

Table 9
Component ablation experiment results on ISIC2018 dataset.

Component	Acc%	F1%	Prec%	Recall%
Local Path	77.12	54.12	60.03	52.77
+ Global Path	79.59	64.32	64.24	64.66
+ Channel&Spatial Attn	80.85	66.26	69.20	64.60
+ IRMLP	81.32	69.53	72.20	68.28
+Shortcut (HiFuse-Tiny)	82.99	72.99	73.67	72.87

HFF block can provide a better fusion of global–local features. The above combination achieves an Acc of 82.99% and an F1 value of 72.99%.

The Table 10 summarizes the performance of each branch of HiFuse-Base under the configuration of fusion of different stages on the ISIC2018 dataset. It can be seen from the ablation results that more layers of HiFuse fusion can bring better results, and all four stages participating in the fusion can achieve the best results. It verifies the necessity of our proposed HiFuse for fusion at all levels, which can fuse global and local features more comprehensively.

4.5. Visual inspection of HiFuse

To further illustrate that our HiFuse model can effectively capture feature information of medical images, we choose the same HiFuse hierarchical structure models ConvNeXt (similar to our proposed local path), Swin-Transformer (similar to our proposed global path), and compare them with our HiFuse in this section. We adopt the method of Grad-CAM [68] to visualize the last layer in the model except for the linear layer and reflect the area of interest in the model in the form of a heat map. Fig. 6 show the Grad-CAM visualization results of some (a) dermoscopy, (b) upper gastrointestinal endoscopy, (c) covid19-CT and (d) esophageal cancer pathology.

As can be seen, ConvNeXt attaches great importance to local features, while Swin-Transformer is better at paying attention to global features. The HiFuse model reflects a higher thermal value in the lesion area and more accurately covers the lesion area. Such observations demonstrate that the HiFuse can better integrate global–local features at different levels and can help the model to learn more discriminative features and pay more attention to the lesion area.

5. Discussion

Capturing multi-scale global–local features has obvious advantages in medical image classification. Our HiFuse can achieve higher accuracy with lower computational complexity, which can be widely used in various medical datasets.

The experimental results showed that HiFuse could achieve the best results in the ISIC2018, Kvasir, Covid-19-CT, and esophageal cancer pathology medical classification datasets, which validated the effectiveness of HiFuse. Compared with HiFuse-Tiny and HiFuse-Small, HiFuse-Base had +2.86% and +2.26% ACC in the ISIC2018 dataset.

Our model also has the limitation that the ACC result of relatively deeper HiFuse-Base is −0.15% and −0.54% and −3.85% than the ACC result of HiFuse-Small in the Kvasir, Covid-19-CT, and esophageal cancer pathology datasets. On the one hand, we believe that smaller datasets should use relatively shallower models, and some methods, such as drop path and data enhancement, are used to improve this phenomenon; on the other hand, there is more room for optimization of the deeper HiFuse.

The HiFuse proposed in this paper has many advantages over other feature extraction methods. It is a three-branch hierarchical feature fusion model that fuses global–local feature representations of different scales through HFF blocks. The experiment did not perform special enhancements (cutmix, mixup) or modify the loss function to enhance

Table 10
Stage fusion ablation experiment results on ISIC2018 dataset.

Method	Stage1	Stage2	Stage3	Stage4	ACC	F1	Prec	Recall	MCC	Kappa	AUC
HiFuse-B				✓	0.8047	0.6479	0.6501	0.6463	0.6183	0.6175	0.9196
			✓	✓	0.8207	0.7085	0.6723	0.6873	0.6435	0.6414	0.9319
		✓	✓	✓	0.8357	0.7396	0.7214	0.7263	0.7263	0.6798	0.9400
	✓	✓	✓	✓	**0.8585**	**0.7457**	**0.7658**	**0.7532**	**0.7282**	**0.7280**	**0.9585**

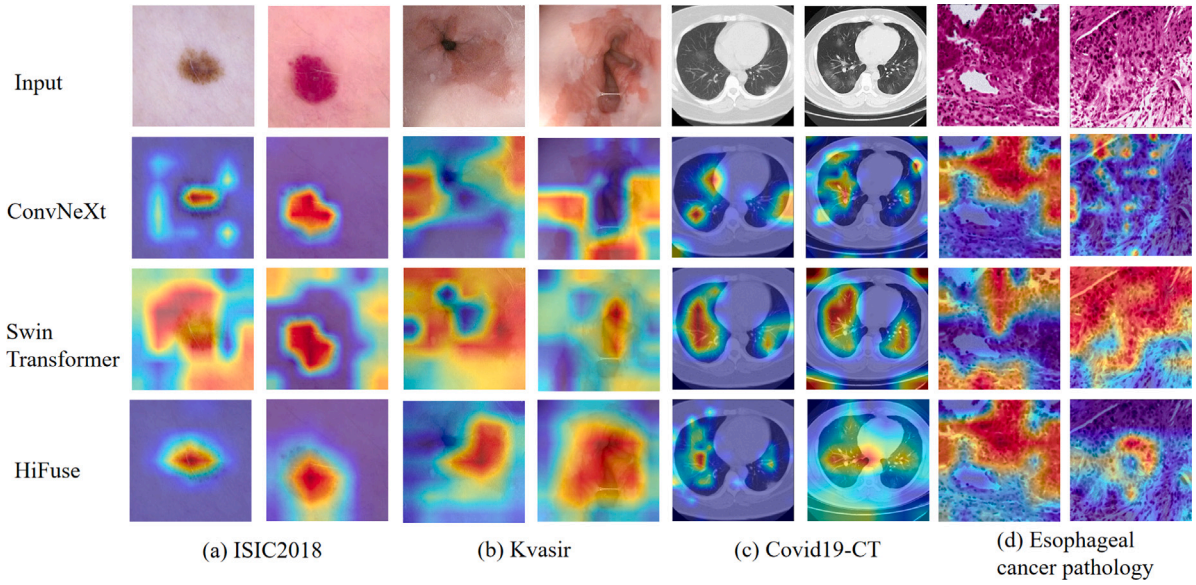

Fig. 6. Grad-CAM visual results of ablation experiments.

the data representation of the dataset. It relies more on the feature extraction capability of the model itself. When combined through HFF blocks at various hierarchical levels, it can integrate the local feature extraction branch (local path) and the global feature branch (global path) and model the global–local multi-scale features in the dataset. The global path and local path do not interact on their respective branches to keep the global–local feature representation of each level mapped from the original image space. It has such a stronger capability to extract global and local features at different levels that it is more suitable for medical datasets.

The HFF block further fuses the features extracted from different branches and sets spatial and channel attention to enhance the expressive capability of the branch and suppress irrelevant features. IRMLP is used to learn the fused features, using depthwise convolution to reduce the computational cost. The shortcut can improve the gradient transfer ability between layers and reduce overfitting to a certain extent, using depthwise convolution, design of global–local feature blocks, W-MSA, and hierarchical structures to reduce computational costs.

Compared with natural images, medical images have diverse characteristics [69], fewer data, and different medical equipments, resulting in limited training data and models that cannot focus well on classification features. Many models performing well in natural image classification tasks are used in the medical field, however, no satisfactory results can be obtained. This paper offers researchers fresh perspectives on fusing global and local features. The hierarchical fusion method in the architecture is designed to be easily extended and upgraded. Our model can be further improved in future research:

1. Under the specific situation of the task, assign the network depth and width of different branches to make the local features and global representation more directional.

2. Design targeted dynamic hierarchical feature selection for different datasets to improve the performance of HiFuse further.

3. Subsequent HiFuse research in downstream tasks such as medical image segmentation and multimodal.

6. Conclusion

In this paper, we propose a novel general-purpose backbone HiFuse for medical image classification, which can efficiently extract local features and global representations through HFF blocks, a three-branch hierarchical multi-scale feature fusion manner, so as to preserve global and local model, and avoid introducing new noises. Our modularly-designed global and local feature blocks have rich scalability and linear computational complexity. Extensive experiments demonstrate the excellent performance of our HiFuse, which can comprehensively mine the features of lesion regions in multi-scale global–local medical image classification tasks. We believe this work can contribute to various downstream tasks in medical imagery.

CRediT authorship contribution statement

Xiangzuo Huo: Conceptualization, Methodology, Software. **Gang Sun:** Data curation, Writing – original draft. **Shengwei Tian:** Visualization. **Yan Wang:** Investigation. **Long Yu:** Investigation. **Jun Long:** Supervision. **Wendong Zhang:** Software, Validation. **Aolun Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62162058, the Xinjiang Uygur Autonomous Region Key Research and Development Project 2021B03001-4, and the Outstanding Doctoral Student Scientific Research Innovation Project XJU2022BS078.

References

[1] S. Koitka, C.M. Friedrich, Traditional feature engineering and deep learning approaches at medical classification task of imageclef 2016, in: CLEF (Working Notes), Citeseer, 2016, pp. 304–317.
[2] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, C. Chang, Deep learning of feature representation with multiple instance learning for medical image analysis, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2014, pp. 1626–1630.
[3] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, J. Tian, Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification, Pattern Recognit. 61 (2017) 663–673.
[4] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Correction: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks, Nature 546 (7660) (2017) 686.
[5] L. Personnaz, I. Guyon, G. Dreyfus, Collective computational properties of neural networks: New learning mechanisms, Phys. Rev. A 34 (5) (1986) 4217.
[6] A. Kumar, J. Kim, D. Lyndon, M. Fulham, D. Feng, An ensemble of fine-tuned convolutional neural networks for medical image classification, IEEE J. Biomed. Health Inform. 21 (1) (2016) 31–40.
[7] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, Z. Zhao, Deep transfer learning for modality classification of medical images, Information 8 (3) (2017) 91.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[10] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.
[11] Y. Xu, Q. Zhang, J. Zhang, D. Tao, Vitae: Vision transformer advanced by exploring intrinsic inductive bias, Adv. Neural Inf. Process. Syst. 34 (2021) 28522–28535.
[12] B. Fu, M. Zhang, J. He, Y. Cao, Y. Guo, R. Wang, StoHisNet: A hybrid multi-classification model with CNN and transformer for gastric pathology images, Comput. Methods Programs Biomed. (2022) 106924.
[13] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 14–24.
[14] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, C. Xu, Cmt: Convolutional neural networks meet vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12175–12185.
[15] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 367–376.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[19] S.H. Baloch, H. Krim, Flexible skew-symmetric shape model for shape representation, classification, and sampling, IEEE Trans. Image Process. 16 (2) (2007) 317–328.
[20] Y. Song, W. Cai, Y. Zhou, D.D. Feng, Feature-based image patch approximation for lung tissue classification, IEEE Trans. Med. Imaging 32 (4) (2013) 797–808.
[21] M. Irfan, M.A. Iftikhar, S. Yasin, U. Draz, T. Ali, S. Hussain, S. Bukhari, A.S. Alwadie, S. Rahman, A. Glowacz, et al., Role of hybrid deep neural networks (HDNNs), computed tomography, and chest X-rays for the detection of COVID-19, Int. J. Environ. Res. Public Health 18 (6) (2021) 3056.
[22] Y.E. Almalki, A. Qayyum, M. Irfan, N. Haider, A. Glowacz, F.M. Alshehri, S.K. Alduraibi, K. Alshamrani, M.A. Alkhalik Basha, A. Alduraibi, et al., A novel method for COVID-19 diagnosis using artificial intelligence in chest X-ray images, in: Healthcare, Vol. 9, No. 5, MDPI, 2021, p. 522.
[23] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, H. Lu, ResGANet: Residual group attention network for medical image classification and segmentation, Med. Image Anal. 76 (2022) 102313.
[24] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 87–110.
[25] S. Jamil, M. Jalil Piran, O.-J. Kwon, A comprehensive survey of transformers for computer vision, Drones 7 (5) (2023) 287.
[26] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. 54 (10s) (2022) 1–41.
[27] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, 2020, arXiv preprint arXiv:2006.03677.
[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[30] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.
[31] J. Beal, E. Kim, E. Tzeng, D.H. Park, A. Zhai, D. Kislyuk, Toward transformer-based object detection, 2020, arXiv preprint arXiv:2012.09958.
[32] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, 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, 2021, pp. 6881–6890.
[33] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, Y. Qian, STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 10990–11003.
[34] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-trained image processing transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12299–12310.
[35] Z. Wan, J. Zhang, D. Chen, J. Liao, High-fidelity pluralistic image completion with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4692–4701.
[36] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 36–46.
[37] X. He, Y. Chen, Z. Lin, Spatial-spectral transformer for hyperspectral image classification, Remote Sens. 13 (3) (2021) 498.
[38] Y. Jiang, S. Chang, Z. Wang, Transgan: Two pure transformers can make one strong gan, and that can scale up, Adv. Neural Inf. Process. Syst. 34 (2021) 14745–14758.
[39] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 109–119.
[40] H. Chen, C. Li, G. Wang, X. Li, M.M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, et al., GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection, Pattern Recognit. 130 (2022) 108827.
[41] J. Wang, S. Tian, L. Yu, Z. Zhou, F. Wang, Y. Wang, HIGF-Net: Hierarchical information-guided fusion network for polyp segmentation based on transformer and convolution feature learning, Comput. Biol. Med. 161 (2023) 107038.
[42] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.
[43] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[44] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[46] R. Yan, F. Ren, J. Li, X. Rao, Z. Lv, C. Zheng, F. Zhang, Nuclei-guided network for breast cancer grading in he-stained pathological images, Sensors 22 (11) (2022) 4061.
[47] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[48] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.

[49] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, Advances in neural information processing systems 27 (2014).
[50] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[51] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.
[52] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R.R. Martin, M.-M. Cheng, S.-M. Hu, Attention mechanisms in computer vision: A survey, Comput. Vis. Media (2022) 1–38.
[53] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal self-attention for local-global interactions in vision transformers, 2021, arXiv preprint arXiv:2107.00641.
[54] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.
[55] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, J. Wang, Hrformer: High-resolution vision transformer for dense predict, Adv. Neural Inf. Process. Syst. 34 (2021) 7281–7293.
[56] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019, arXiv preprint arXiv:1902.03368.
[57] K. Pogorelov, K.R. Randel, C. Griwodz, S.L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P.T. Schmidt, et al., Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, pp. 164–169.
[58] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, P. Xie, Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans, Cold Spring Harbor Laboratory Press, 2020, medrxiv.
[59] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A.M. Schlitter, I. Esposito, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, IEEE Trans. Med. Imaging 35 (8) (2016) 1962–1971.
[60] MM.C.V. Contributors, MMCV: OpenMMLab computer vision foundation, 2018, https://github.com/open-mmlab/mmcv.
[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[62] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.
[63] J. Min, Y. Zhao, C. Luo, M. Cho, Peripheral vision transformer, Adv. Neural Inf. Process. Syst. 35 (2022) 32097–32111.
[64] J. Yang, C. Li, X. Dai, J. Gao, Focal modulation networks, Adv. Neural Inf. Process. Syst. 35 (2022) 4203–4217.
[65] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, Y. Qiao, Uniformer: Unifying convolution and self-attention for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[66] L. Zhu, X. Wang, Z. Ke, W. Zhang, R.W. Lau, Biformer: Vision transformer with bi-level routing attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10323–10333.
[67] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers i& amp; distillation through attention, in: International Conference on Machine Learning, Vol. 139, 2021, pp. 10347–10357.
[68] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[69] S.K. Zhou, H. Greenspan, C. Davatzikos, J.S. Duncan, B. Van Ginneken, A. Madabhushi, J.L. Prince, D. Rueckert, R.M. Summers, A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, Proc. IEEE 109 (5) (2021) 820–838.