

Omid Nejadi Manzari^{a,*}, Hamid Ahmadabadi^a, Hossein Kashiani^b, Shahriar B. Shokouhi^a and Ahmad Ayatollahi^a^aSchool of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran^bLane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA

ARTICLE INFO

Keywords:

Medical image classification
 Adversarial attack
 Adversarial robustness
 Vision Transformer

ABSTRACT

Convolutional Neural Networks (CNNs) have advanced existing medical systems for automatic disease diagnosis. However, there are still concerns about the reliability of deep medical diagnosis systems against the potential threats of adversarial attacks since inaccurate diagnosis could lead to disastrous consequences in the safety realm. In this study, we propose a highly robust yet efficient CNN-Transformer hybrid model which is equipped with the locality of CNNs as well as the global connectivity of vision Transformers. To mitigate the high quadratic complexity of the self-attention mechanism while jointly attending to information in various representation subspaces, we construct our attention mechanism by means of an efficient convolution operation. Moreover, to alleviate the fragility of our Transformer model against adversarial attacks, we attempt to learn smoother decision boundaries. To this end, we augment the shape information of an image in the high-level feature space by permuting the feature mean and variance within mini-batches. With less computational complexity, our proposed hybrid model demonstrates its high robustness and generalization ability compared to the state-of-the-art studies on a large-scale collection of standardized MedMNIST-2D datasets.

1. Introduction

Medical image classification is a critical step in medical image analysis that uses different factors such as clinical information or imaging modalities to differentiate across medical images. A dependable medical image classification may help clinicians evaluate medical images quickly and with less error. The healthcare industry has significantly benefited from recent Convolution Neural Networks (CNNs) advancements. Such advancements have prompted much research into the use of computer-aided diagnostic systems [1–4] based on artificial intelligence in clinical settings. CNNs are able to learn robust discriminative representation from vast volumes of medical data to generate accurate diagnostic performance in medical fields. They validate their satisfactory prediction capabilities and obtain comparable performance as clinicians.

However, the locality bias of CNNs makes it hard for them to learn long-range dependencies in visual data. The texture, shape, and size of many organs vary widely across people, making it difficult to correctly analyze medical data [5, 6]. As such, it is important to extract robust feature representation which can model long-range dependencies in different domains for medical image analysis. Recently, the Transformer architectures have adopted the self-attention mechanisms to model the long-range dependencies between input images and have achieved promising results. Different studies demonstrate their performance superiority compared to CNN architectures [7, 8]. However, a sizable amount of training data is crucial to their success. The construction of a large-scale dataset needs a significant amount of time and resources.

*Corresponding author.

✉ omid_nejaty@alumni.iust.ac.ir (O.N. Manzari)
 ORCID(s):

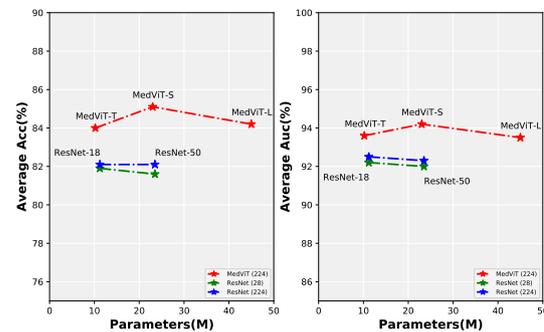


Figure 1: Comparison between MedViTs and the baseline ResNets, in terms of average ACC-Parameters and average AUC-Parameters trade-off over all 2D datasets.

Regarding the medical field, radiologist experts must manually annotate and verify medical data, which is costly and time-consuming. While the Transformer architecture mitigates the shortcomings of CNNs, its computational complexity grows quadratically with spatial or embedding dimensions, therefore making it infeasible for most image restoration tasks involving high-resolution images. That is to say, the Transformer architectures address the long-range dependency modeling in CNNs, yet their computational complexity increases quadratically with the spatial dimension [9]. As a result, they cannot be used in realistic clinical settings. Moreover, the state-of-the-art studies assume that training and test data are identically distributed. Consequently, on out-of-domain target domains, they often suffer significant performance drops. The domain shift is more pronounced in healthcare areas since medical images can be captured by different devices at various sites. Consequently,

due to different scanners and imaging protocols, their data distribution can greatly vary. In addition, variations in epidemiology at different sites could impact the distribution of ground truth labels between various populations [10, 11].

In this study, we aim to address the above-mentioned challenges and propose a generalized Transformer architecture for medical image analysis. While recent studies in medical image analysis work specifically on the pre-determined medical test sets, our proposed model generalizes to a wide range of medical domains such as CT, X-ray, ultrasound, and OCT domains. To this end, we follow the hierarchical hybrid architecture equipped with a patch embedding layer and a series of convolution and Transformer blocks in each stage with efficient computational complexity. Inspired by recent advances, we cater out Transformer architecture in such a way that each stage consists of two efficient phases to long-term dependencies and model short-term in visual data. In the first phase, we leverage a multi-head convolutional attention block to learn affinity between different tokens in representation subspaces for effective local representation learning. The attention block alleviates the high computational attention-based token mixer in conventional Transformer architectures, thereby improving inference speed.

Compared to the conventional Transformer architectures that suffice to incorporate locality bias into the lower layers of Transformer architectures through tokenization and self-attention components (CvT [12], Co-Scale Conv-Attentional [13], CMT [14]). We also propose a local feed-forward network (LFFN) that encodes the local dependencies between nearby pixels in feed-forward components of Transformer architectures in all stages. For this objective, a depth-wise convolution is applied to the reshaped 2D feature map. While recent Transformer-based studies have demonstrated a high capacity to learn long-range dependencies between visual data, they fail to encode high-frequency context in visual data. To mitigate this issue, in the second phase of our proposed architecture, we first encode low- and high-frequency feature representation separately with an efficient multi-head self-attention block and a multi-head convolutional attention block, respectively. Then, the computed feature representations are fused and fed to the LFFN to enhance global and local modeling capacity further. As depicted in Figure 1, our model shows great superiority in terms of accuracy-complexity against CNNs.

The adversarial attack is a serious security risk for deep neural networks because it could trick the trained models into making incorrect predictions via small, undetectable perturbations. When it comes to healthcare, an adversarial attack could also pose severe security concerns [15]. The nature of medical data could provide an opportunity for a higher attack success rate with imperceptibility. In this paper, to enhance the adversarial robustness of our proposed Transformer model, we make our model focus more on global structure features (such as shape and style) rather than texture information. Several works [16, 17] have found

that neural networks tend to rely upon texture information for making predictions, which consequently makes them vulnerable to out-of-distribution samples. Motivated by these studies, we try to encourage our model to rely more on global structure features rather than texture information to boost the generalization performance as well as adversarial robustness. With this objective in mind, we extract the mean and variance of training instances across channel dimensions in feature space and interpolate them with each other. It is worth pointing out that the decision boundary is often sharp, and there is a significant portion of the hidden representation space that is associated with high-confidence predictions [18, 19]. With the proposed interpolation, we can explore new useful regions of the feature space, which are mainly relevant to global structure features. This ultimately would enable us to learn smoother decision boundaries, which are beneficial for adversarial robustness and generalization performance.

2. Related Works

Convolutional networks. Convolutional neural networks (CNNs) have witnessed extraordinary contributions to the vast fields of computer vision in recent years due to their ability to extract deep discriminative features. ResNet [20] introduced residual connections to CNN and mitigated the vanishing gradient problem, which ensures the model builds deeper to capture high-level features for image classification. MobileNets [21] use pointwise convolutions and depthwise separable convolutions to enhance CNN efficiency. In DenseNet [22], skip connections were used between each two layers, and summation was replaced with concatenation for the dense connections of feature maps. ConvNext [23] re-introduces core designs of Vision Transformers and employs 7×7 depthwise convolutions to design robust CNN architecture, which can achieve comparable results with Transformers. ShuffleNet [24] performs the channel shuffle operation to fuse separated channel information using group convolution.

Vision Transformers. Since original Transformer architecture achieved remarkable results in natural language processing many attempts have been made to use Transformer architecture to vision tasks like image classification [7], semantic segmentation [25], and object detection [26]. In particular, the Vision Transformer (ViT) of Dosovitskiy et al. [7] shows that pure Transformer-based can also achieve promising result on the image classification task. ViT splits the image into patches (a.k.a., tokens) and applies transformer layers to model the global relation among these patches for classification. T2T-ViT [27] mainly improves tokenization in ViT by delicately generating tokens in a soft split manner, which recursively aggregates neighboring tokens into one token to enrich local structure modeling. Swin Transformer [28] performs self-attention in a local window with the shifted window scheme to alternately model in-window and cross-window connection. PiT [29] follows a similar pyramid structure as

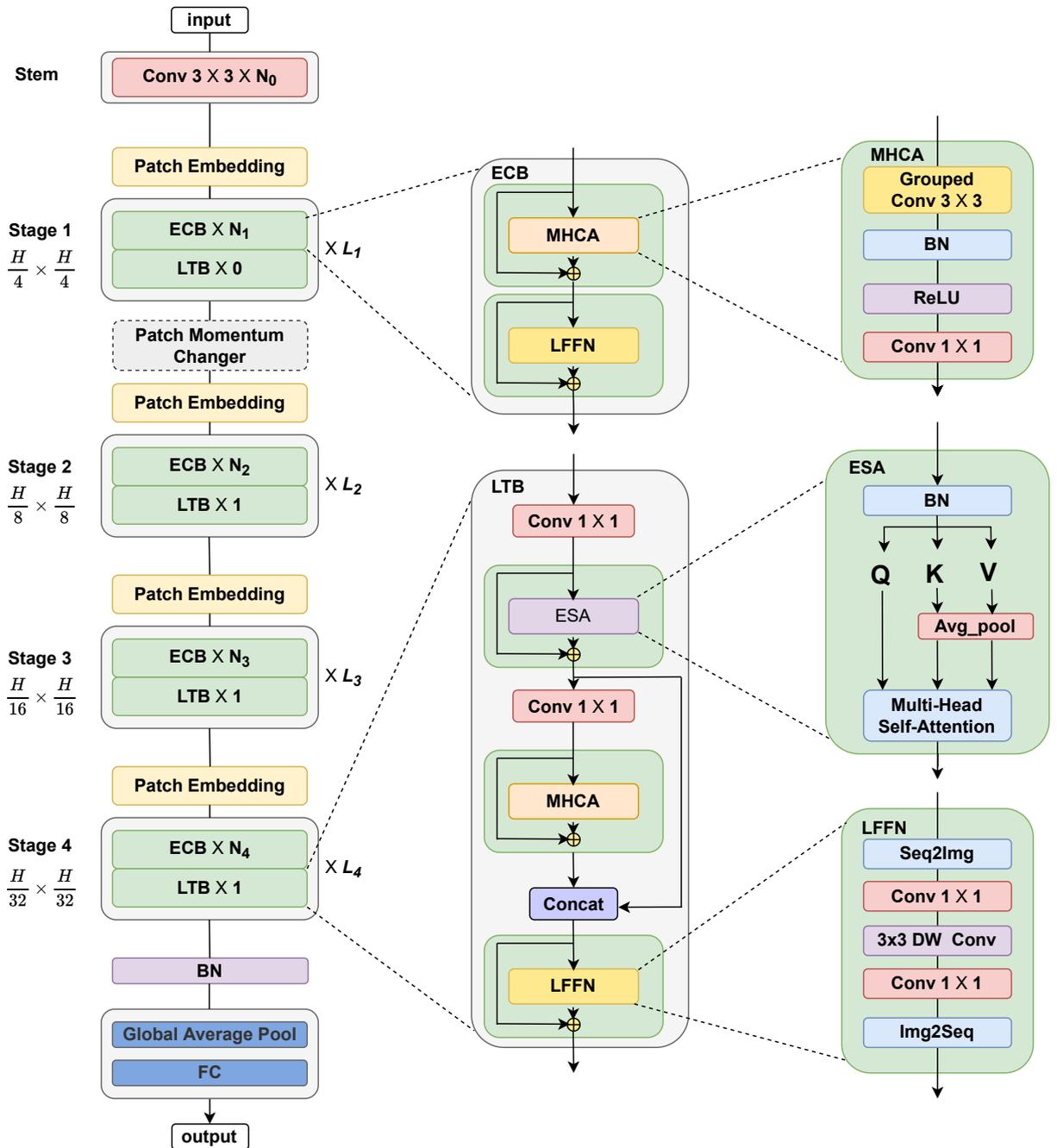


Figure 2: Overall architecture of the proposed Medical Vision Transformer (MedViT).

CNNs, which produces various feature maps through spatial dimension reduction based on the pooling structure of a convolutional layer. Nowadays, researchers are specifically interested in efficient methods, including pyramidal designs, training strategies, efficient self-attention, etc.

Hybrid Models. Recent works show that designing a hybrid architecture of transformer and convolution layers helps the model to combine the advantages of both architectures. BoTNet [30] uses a slightly-modified self-attention in the last three blocks of ResNet. CMT [14] block contains depthwise convolution layers based local perception unit and a lightweight transformer block. The CvT [12] inserts pointwise and depthwise convolution before self-attention.

LeViT [31] uses the convolutional stem to replace the patch embedding block and achieves fast inference image classification. The MobileViT [32] introduces a lightweight vision transformer by combining Transformer blocks with the MobileNetV2 [33] block in series. Mobile-Former [34] takes a bidirectional bridge between CNN and transformer to leverage the advantage of global and local concepts.

Robustness Study. Due to the nature of convolutional neural networks that rely on low-level features, their assumptions are generally vulnerable to adversarial examples. There are numerous studies on improving the adversarial robustness of CNNs that aim to strengthen it in various approaches. These include carefully designed model [35,

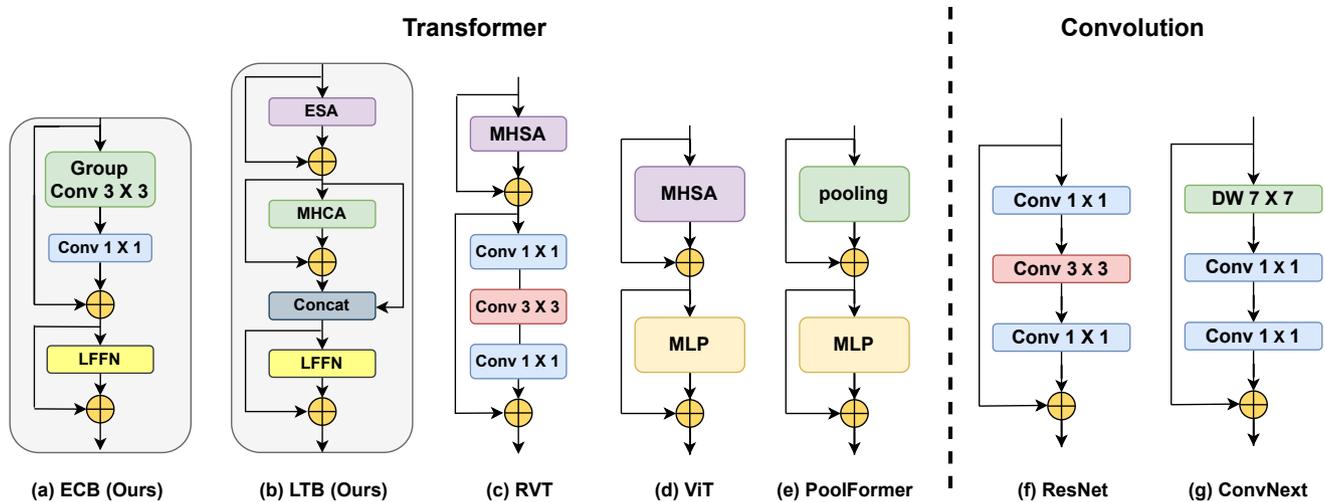


Figure 3: Comparison of different core blocks. Convolution-based that includes the ResNet, ConvNext, and Transformer-based that includes RVT, PoolFormer, ViT, and MedViT (Ours).

36], strong data augmentation [37, 38], searched network architecture [39, 40], improved training strategy [41–43], pruning [44] of the weights, and quantization [45], activation functions [46] or better pooling [47, 48]. Although the methods mentioned earlier perform well on CNNs, there is no evidence that they also improve the adversarial robustness of ViTs.

Following the success of Transformers and their variants in various computer vision tasks, several studies are attempting to examine the robustness of Transformers. Early research conduct the adversarial robustness of Transformers on image classification tasks and compare their vulnerability against the MLP and CNN baselines. The experimental results illustrate that Transformers are more adversarially robust than CNNs [49]. Additionally, the adversarial transferability between CNNs and Transformers is unexpectedly low[50]. Furthermore, the robustness study [51] of ViTs is extended to the natural distribution shift and common image corruption, which demonstrate the superiority of ViTs over CNNs in the robustness benchmark. Although several studies have challenged the adversarial robustness without carefully designing architecture, in this paper, we do not make a simple comparison of adversarial robustness between CNNs and ViTs, but take a step further by designing a robust hybrid architecture family of MedViTs. Based on the architecture of Transformer, we introduce a novel Augmentation technique to further reduce the fragility of Transformer models.

3. Method

We first give a brief overview of the proposed MedViT in this section. Then, We describe the main body designs within MedViT, which include the Efficient Convolution Block (ECB), Local Transformer Block (LTB), and Transformer Augmentation Block (TAB). In addition, we provide different model sizes for the proposed architecture.

3.1. Overview

MedViT aims to combine the convolution block and transformer block in a novel approach to achieve a robust hybrid architecture for medical image classification. As shown in Figure 2, MedViT is composed of a patch embedding layer, transformer blocks and a series of stacked convolution in each stage, which follows the hierarchical pyramid architecture traditionally. The spatial resolution will be gradually reduced with a total of $32\times$ ratio by $[4\times, 2\times, 2\times, \text{and } 2\times]$ while the channel dimension will be doubled after convolution blocks in each stage. Our purpose in this section is first to explore the core blocks are responsible for embedding multi-scale context and respectively develop robust LTB and ECB to effectively capture long-term and short-term dependencies in input data. LTB also performs the fusion of local and global features, thereby enhancing modeling capabilities. Also we study how to integrate blocks of convolution and transformer technically. Lastly, to further improve the performance and adversarial robustness, we propose a novel Patch Momentum Changer (PMC) data augmentation technique to train our models.

3.2. Efficient Convolution Block

We begin by discussing some traditional core blocks of transformer and convolution network, as illustrated in Figure 3. To show the effectiveness of the proposed ECB and its superiority over previous methods. ResNet [20] introduced skip connection and Residual block, which has dominated a wide range of tasks in visual recognition for a long time due to its compatible features and inherent inductive biases for the realistic deployment scenario. Unfortunately, the performance of the Residual block is not satisfactory compared to the Transformer block. The ConvNext block [23] constructed from the Residual block by following the designs of the Transformer block without self-attention. Although the ConvNext block enhances network performance to some extent, inefficient components make the model hard to capture high-level structures, such as 7×7

depth convolution, GELU and LayerNorm. To overcome this, the transformer block has been proposed to capture high-level structures. Transformers have achieved excellent results in various computer vision tasks, and their inherent superiority is jointly conferred by the attention-based token mixing operation [28]. However, these methods merely focus on the model complexity and the standard accuracy. The results demonstrate that the models are vulnerable to adversarial attacks [52, 53], which are intolerable in clinically relevant medical use cases.

In addition, long-range dependencies are crucial for medical images because the background in medical images is generally scattered [54], thereby learning long-range dependencies between the pixels corresponding to the background can help the network in preventing misclassification [55]. It can be noted that there is still scope for improvement in capturing global context as the shortcoming of prior methods, which do not focus on this aspect for medical image classification tasks. To address the adversarial robustness and accurate medical classification, we introduce an Efficient Convolution Block (ECB) that achieves outstanding performance as a transformer-based block while retaining the deployment advantage of the Residual block. As illustrated in Figure 3 (a), The ECB follows the hybrid architecture, which has been confirmed as necessary for utilizing the multi-scale information. Meanwhile, an effective attention-based token mixing module is equally important. We design a Locally Feed Forward Network (LFFN) as an efficient way of introducing locality into the network with depth-wise convolution and a Multi-Head Convolutional Attention (MHCA) as an effective token mixer. Inspired by Robust Vision Transformer [56] that analyzed the effect of each component of Transformers in the robustness, we build ECB by combining LFFN and MHCA block in the robust paradigm. The proposed ECB can be formulated as follows:

$$\begin{aligned} \tilde{z}^l &= \text{MHCA}(z^{l-1}) + z^{l-1} \\ z^l &= \text{LFFN}(\tilde{z}^l) + \tilde{z}^l \end{aligned} \quad (1)$$

where z^{l-1} denotes the input from the $l - 1$ block, \tilde{z}^l and z^l are the outputs of MHCA and the l ECB. We will introduce LFFN in detail in the next section.

3.2.1. Locally Feed-Forward Network

The feed-forward network, which is applied position-wise to a sequence of tokens \mathbf{Z}^r , can be precisely represented by rearranging the sequence of tokens into a 2D lattice, as shown in Figure 4 (c). As a result, the reshaped features are represented as follows:

$$\mathbf{Z}^r = \text{Seq2Img}(\mathbf{Z}), \mathbf{Z}^r \in \mathbb{R}^{h \times w \times d} \quad (2)$$

where $h = H/p$ and $w = W/p$. *Seq2IMG* takes a sequence and converts it into a feature map that can be visualized. The tokens are placed at pixel locations on the feature map, and each token corresponds to one pixel.

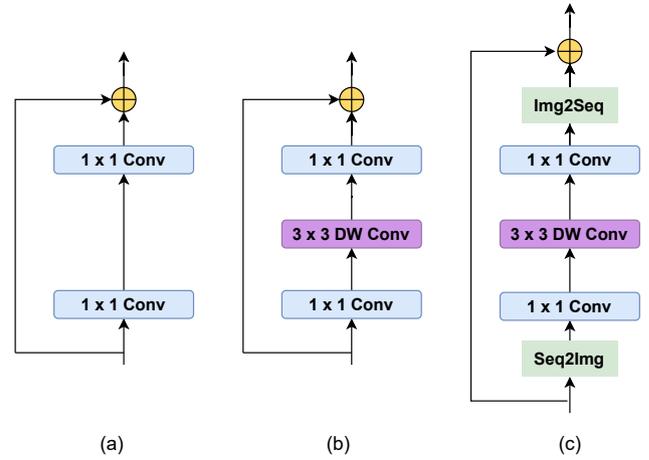


Figure 4: Comparison between the feed-forward network in vision transformers, (a) convolutional feed-forward network. (b) inverted residual block. (c) Our finally utilized network that brings efficient local mechanism into the transformer.

Through this perspective, it is possible to introduce locality into the network by recovering the proximity between tokens. The fully-connected layers could be replaced by 1×1 convolution layers, i.e.

$$\begin{aligned} \mathbf{Y}^r &= f(\mathbf{Z}^r \otimes \mathbf{W}_1^r) \otimes \mathbf{W}_2^r \\ \mathbf{Y} &= \text{Img2Seq}(\mathbf{Y}^r) \end{aligned} \quad (3)$$

where $\mathbf{W}_1^r \in \mathbb{R}^{d \times \gamma d \times 1 \times 1}$ and $\mathbf{W}_2^r \in \mathbb{R}^{\gamma d \times d \times 1 \times 1}$ are reshaped from \mathbf{W}_1 and \mathbf{W}_2 and denote the kernels of convolutional layers. With *Img2Seq*, the image feature map is converted back into a token sequence, which is then used by the next self-attention layer by transforming it into the fused token.

3.3. Local Transformer Block

While the local representation has been effectively learned through the ECB, capturing global information is urgent and needs to be addressed in this block. It is well known that transformer blocks have the capability to capture low-frequency signals, which are very useful for capturing global information (e.g., global shapes and structures). However, There have been a few related studies [57] that have demonstrated that transformer blocks have a tendency to deteriorate high-frequency information, such as information about the local texture of objects, to some extent. It is essential that signals in different frequency segments are fused in order to extract essential and distinct features in the computer vision system [58].

In response to these observations, we have developed the Local Transformer Block (LTB) in order to capture multi-frequency signals in a lightweight mechanism with high efficiency. Moreover, LTB works as an effective multi-frequency signal mixer, thus enhancing the overall modeling capability of the network. As shown in Figure 3 (b), LTB first captures low-frequency signals by utilizing an Efficient

Table 1

Detailed configurations of MedViT variants. C and S denotes number of channels and stride of convolution for each stage.

Stages	Output size	Layers	MedViT-T	MedViT-S	MedViT-L
Stem	$\frac{H}{4} \times \frac{W}{4}$	Convolution Layers		Conv 3 × 3, C = 64, S = 2	
				Conv 3 × 3, C = 32, S = 1	
				Conv 3 × 3, C = 64, S = 1	
				Conv 3 × 3, C = 64, S = 2	
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding		Conv 1 × 1, C = 96	
		MedViT Block		[ECB × 3, 96] × 1	
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding		Avg_pool, S = 2	
		MedViT Block		Conv 1 × 1, C = 192	
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding		Avg_pool, S = 2	
		MedViT Block	[ECB × 4, 384] LTB × 1, 512] × 2	[ECB × 4, 384] LTB × 1, 512] × 4	[ECB × 4, 384] LTB × 1, 512] × 6
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding		Conv 1 × 1, C = 768	
		MedViT Block		[ECB × 2, 768] LTB × 1, 1024] × 1	

Self Attention (ESA), which can be formulated as follows:

$$ESA(x) = \text{Concat} (SA_1(x_1), SA_2(x_2), \dots, SA_h(x_h)) W^O$$

$$SA(X) = \text{Attention} (X \cdot W^Q, P_s(X \cdot W^K), P_s(X \cdot W^V)) \quad (4)$$

where $X = [x_1, x_2, \dots, x_h]$ denotes to divide the input feature X into multi-head form in channel dimension. W^O is an output projection layer and h is the number of heads. In order to reduce the spatial resolution of self-attention, SA was derived from linear SRA [59]. Attention calculates W^Q, W^K, W^V linear layers in standard attention form as $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d})V$, in which d is the transformer hidden dimension. The P_s operation involves an avg-pool operation with a stride s parameter for downsampling the spatial dimensions before the attention operation is applied to reduce the computation cost. Furthermore, we observe that the number of channels in the ESA module is also a major determinant of the time consumption of the module. With the help of point-wise convolutions, LTb enables further acceleration of inference by reducing the dimension of the channel before it passes to the ESA module. In order to reduce the number of channels, a shrinking ratio r is introduced. Additionally, the ESA module also utilizes Batch Normalization to make the module's deployment extremely efficient.

It is significant to note that LTb has a multi-frequency configuration that is designed to function in conjunction with ESA and MHCA modules. Following that, we design a new attention mechanism that is based on efficient convolutional operations for improving the efficiency of the LTb. Inspired by the effective multi-head design in MHSA [36], we build our convolutional attention (CA) with multi-head paradigm, which jointly attends to information from different representation subspaces at different positions for effective local representation learning. The

proposed MHCA can be formulated as follows:

$$MHCA(x) = \text{Concat} (CA_1(x_1), CA_2(x_2), \dots, CA_h(x_h)) W^O$$

$$CA(X) = (W \cdot T_{\{i,j\}}), \text{ where } T_{\{i,j\}} \in X \quad (5)$$

where MHCA captures information from h parallel representation subspaces and CA is single-head convolutional attention. W is trainable parameter and $T_{\{i,j\}}$ are adjacent tokens in input feature X . The CA is calculated by the inner product operation of displaced vectors between adjacent tokens $T_{\{i,j\}}$ and trainable parameter W . Since the multi-head self-attention (MHSA) in Transformers could capture the global context, we propose the CA from the MHCA, which can learn affinity between different tokens in the local receptive. Notably, our implementation of MHCA involves a point-wise convolution and a group convolution (multi-head convolution), as shown in Figure 3 (a).

the output features of the MHCA and the ESA are concatenated to produce a mix of high-low features. As a final step, an MLP layer is borrowed at the end of the process in order to extract the essential and distinct features. In brief, the implementation of the LTb can be summarized as follows:

$$\bar{z}^l = \text{Proj}(z^{l-1})$$

$$\bar{z}^l = \text{ESA}(\bar{z}^l) + \bar{z}^l$$

$$\hat{z}^l = \text{Proj}(\bar{z}^l)$$

$$\hat{z}^l = \text{MHCA}(\hat{z}^l) + \hat{z}^l \quad (6)$$

$$\hat{z}^l = \text{Concat}(\hat{z}^l, \bar{z}^l)$$

$$z^l = \text{LFFN}(\hat{z}^l) + \hat{z}^l$$

where z^l is the output of LTb from the l -th block, and \bar{z}^l, \hat{z}^l denote the output of MHCA and ESA, respectively. Proj refers to the point-wise convolution layer associated with the project channel. In order to provide efficient norm and activation layers for LTb, the BN and the ReLU are uniformly adopted instead of the LN and the GELU as the efficient norm and activation layers. A major advantage of the LTb over traditional transformer blocks is its ability to capture and mix multi-frequency information in such a lightweight mechanism, so the performance of the model is greatly enhanced.

3.4. Transformer Augmentation Block

Image augmentation techniques apply geometric transformation functions such as rotating, cropping, and flipping or color space transformation functions such as edge enhancement, grayscale transformations, and color jittering on an input image. Data augmentation is an important strategy for ViTs because they suffer from data scarcity when trained on relatively small-size datasets, while is a data-space solution to the problem of limited data can be solved by strong data augmentation [60]. Moreover, a rich data augmentation also helps with robustness and generalization, which has been verified in previous works [17, 61, 62].

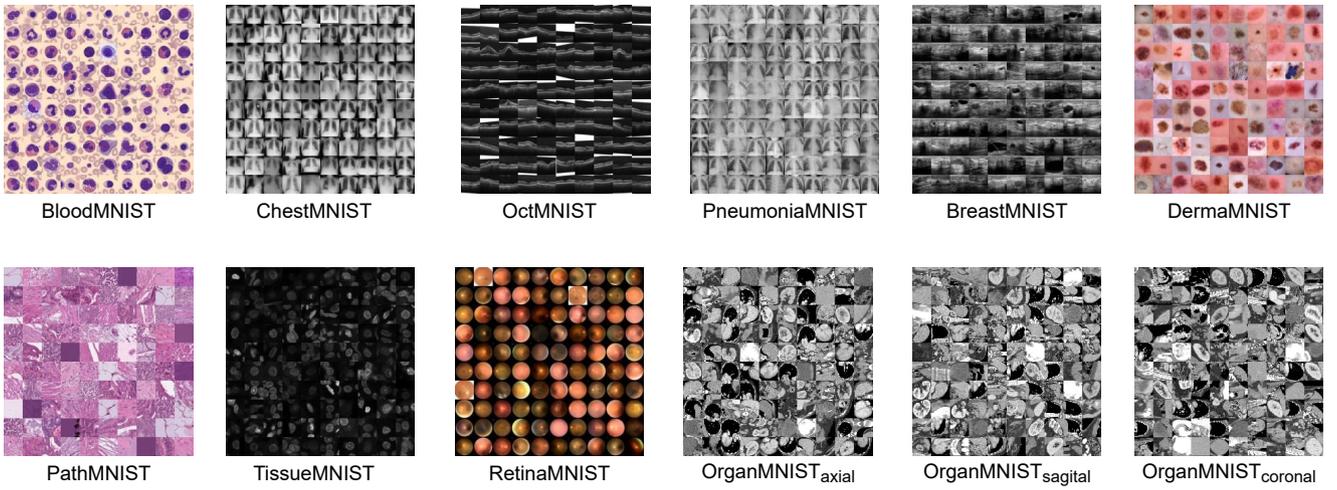


Figure 5: MedMNIST-2D Classification. MedMNIST is a collection of 12 pre-processed medical image datasets. It is designed to be educational, standardized, diverse and lightweight, which could be used as a general classification benchmark in medical image analysis.

In order to improve the diversity of the augmented training data, we introduce Patch Momentum Changer (PMC) augmentation for ViTs, which blends feature normalization with data augmentation at training time for a pair of images at token level. Our motivation stems from the fact that all layers of ViTs have global receptive fields, so they are also concerned with the local relationships that exist between the tokens. We believe that the traditional augmentations, which randomly transform the whole image in order to enlarge the data, are sufficient for providing global context. However, for ViTs that naturally capture global receptive fields, conventional augmentations are less beneficial. In order to increase the interactions between the tokens, PMC laterally fuses intermediate feature maps and targets across two training samples. It mixes two very different components, the feature moments of one instance are combined with the normalized features of another at token level. This asymmetric composition in feature space helps Transformer to improve robustness and generalization when predicting medical image datasets.

At each stage, After feeding the word-level features \hat{Z}^l into the Locally Feed-Forward Network, PMC could take as input the feature representation Z^l , which is a 2D tensor. Similar to Cutmix and Mixup, features of two random training samples are fused with their labels, while performing the feature normalization. Specifically, PMC combine the normalized feature map of one sample with the feature moments of another. This nonsymmetric combination in word-level feature aims to create robust targets and smooth out the decision boundary of the trained classifier. To normalize features at different stages inside the MedViT model, function F is defined. This function takes the word-level features Z_i^l of the i -th input x_i at stage l of MedViT model and generates three outputs which include: the first-moment μ_i , the second moment σ_i , and the normalized word-level features $|Z_i^l|$, as follows:

$$F(Z_i^l) = (\mu_i^l, \sigma_i^l, |Z_i^l|). \quad (7)$$

Function F calculates the value of the first and second momentum after feeding the word-level feature Z_i^l through the LFFN module. This operation relatively resembles PONO function [16] in the realm of CNNs. To employ MedViT model, we randomly select two different images x_A and x_B . The operation could apply at each stage of the model, but it is more effective at the first stage. Consequently, we drop the l superscript for notational simplicity. Augmented features are generated from normalized word-level features of the first image ($x_A : F(Z_A) = (\mu_A, \sigma_A, |Z_A|)$) that are combined with the moments of the second image ($x_B : F(Z_B) = (\mu_B, \sigma_B, |Z_B|)$) as follows:

$$Z_A^{(B)} = \sigma_B \frac{|Z_A| - \mu_A}{\sigma_A} + \mu_B, \quad (8)$$

where $Z_A^{(B)}$ are augmented features and $|Z_A|, \mu_A, \sigma_A$ are the normalized word-level features, the first-moment, and the second moment of image A . In addition, μ_B and σ_B are the first and second moments of image B . The model continues the forward pass from stage l until the output using these features $Z_A^{(B)}$. The lost function is modified to force the model to pay attention to injected features of image x_B . The mixed new loss function would be created as follows:

$$\lambda \cdot (Z_A^{(B)}, y_A) + (1 - \lambda) \cdot (Z_A^{(B)}, y_B), \quad (9)$$

where $\lambda \in (0, 1)$ is a fixed variable for setting the combination of the features and the moments. Also (y_A, y_B) are labels of images, which are combined together for the final loss.

PMC is performed entirely at the feature level inside the transformer vision network and can be readily combined with other augmentation methods that operate on

Table 2

Overview of MedMNIST v2 [63] dataset. MedMNIST2D consists of 12 biomedical datasets of 2D images. Some of the notations used in datasets include OR: Ordinal Regression. MC: Multi-Class. ML: Multi-Label. BC: Binary-Class.

Name	Data Modality	Task (# Classes / Labels)	# Samples	# Training / Validation / Test
MedMNIST2D				
ChestMNIST	Chest X-Ray	ML (14) BC (2)	112,120	78,468 / 11,219 / 22,433
PathMNIST	Colon Pathology	MC (9)	107,180	89,996 / 10,004 / 7,180
OCTMNIST	Retinal OCT	MC (4)	109,309	97,477 / 10,832 / 1,000
DermaMNIST	Dermatoscope	MC (7)	10,015	7,007 / 1,003 / 2,005
RetinaMNIST	Fundus Camera	OR (5)	1,600	1,080 / 120 / 400
PneumoniaMNIST	Chest X-Ray	BC (2)	5,856	4,708 / 524 / 624
BreastMNIST	Breast Ultrasound	BC (2)	780	546 / 78 / 156
TissueMNIST	Kidney Cortex Microscope	MC (8)	236,386	165,466 / 23,640 / 47,280
BloodMNIST	Blood Cell Microscope	MC (8)	17,092	11,959 / 1,712 / 3,421
OrganAMNIST	Abdominal CT	MC (11)	58,850	34,581 / 6,491 / 17,778
OrganCMNIST	Abdominal CT	MC (11)	23,660	13,000 / 2,392 / 8,268
OrganSMNIST	Abdominal CT	MC (11)	25,221	13,940 / 2,452 / 8,829

the raw input (pixels or words). We explicitly encourage the transformer to encode better long-range dependency to correctly classify the image with the feature of another image combined inside. We show that our approach can lead to consistent accuracy gain when used in MedViT, and also enhances the adversarial robustness of the transformers. Besides, we evaluate the efficacy of PMC thoroughly across several datasets.

4. Experiments

4.1. Datasets

MedMNIST datasets include a set of 12 pre-processed datasets that include CT, X-ray, ultrasound and OCT images. These datasets are used in various classification tasks including multi-label, ordinal, multi-class, regression, and binary. The size of the data in this collection varies from at least 100 to more than 100,000. As shown in Table 2, the diversity of these datasets has created a favorable criterion for classification tasks. The pre-processing and split of the datasets into training, validation and test subsets have been done according to [63].

PathMNIST is adapted from a dataset based on kather’s work [64]. This dataset contains 100,000 image patches that are manually divided into 9 different classes. It is also adapted from another dataset that contains 7180 non-overlapping image patches in the classes of fat loss, background, debris, lymphocytes, mucosa, smooth muscle, normal colon mucosa, cancer-related stroma, and epithelium.

ChestMNIST is adapted from a dataset consisting of chest x-ray images [65]. The dataset consists of 112,120 frontal X-ray images from a total of 32,717 patients. This dataset contains 14 different classes of diseases, which is a multi-class dataset in the MEDMNIST collection. We used benchmark standards to split and resize data.

DermaMNIST is based on HAM10000 [66], a large collection of multi-source dermatoscopic images of common pigmented skin lesions. The dataset consists of 10015 dermatoscopic images of a size of 450 × 600. It consists of 7 diagnostic categories as follows: Melanocytic Nevi

(NV), Melanoma (MEL), Basal Cell Carcinoma (BCC), and Intra-Epithelial Carcinoma (AKIEC), Actinic Keratosis, B-nign Keratosis (BKL), Vascular lesions (VASC), Dermatofibroma (DF). All formulated as a multi-class classification task.

OCTMNIST is built on the back of a prior set [67] of 109309 valid optical coherence tomography (OCT) images that were collected for retinal diseases. 4 types are involved, leading to a multi-class classification task.

PneumoniaMNIST is adapted from a prior dataset [68]. This dataset consists of 5856 pediatric chest X-ray images on just two classes. The task is to categorize pneumonia into two binary classes, pneumonia and normal.

RetinaMNIST is based on DeepDRiD (Deep Diabetic Retinopathy) [69], the dataset provides 628 patients data including 1600 retina fundus images. The task is ordinal regression for 5-level grading of diabetic retinopathy severity.

TissueMNIST is adapted from the Broad Bioimage Benchmark Collection [70]. This dataset is categorized into 8 classes of human kidney cortex cells, which contains 236,386 segmented images from different reference tissue specimens.

BloodMNIST is adapted from a prior blood collection [71]. The dataset is categorized into 8 classes and contains a total of 17,092 normal blood cell images.

BreastMNIST is based on a dataset [72] of 780 breast ultrasound images. It is categorized into 3 classes: benign, malignant, and normal. Because low-resolution images are used, the task are simplified into binary classification by combing normal and benign as positive, and classify them against malignant as negative.

OrganMNIST {Axial, Coronal, Sagittal} is taken from 3D computed tomography (CT) images by the Liver Tumor Segmentation Benchmark (LiTS) [71]. As a way to get the organ labels, bounding-box annotations of 11 body organs from another study have been used [70]. As a result of translating the Hounsfield-Unit (HU) of the 3D images to greyscale and with an abdominal window, it is then possible to produce 2D images by selecting slices in the axial,

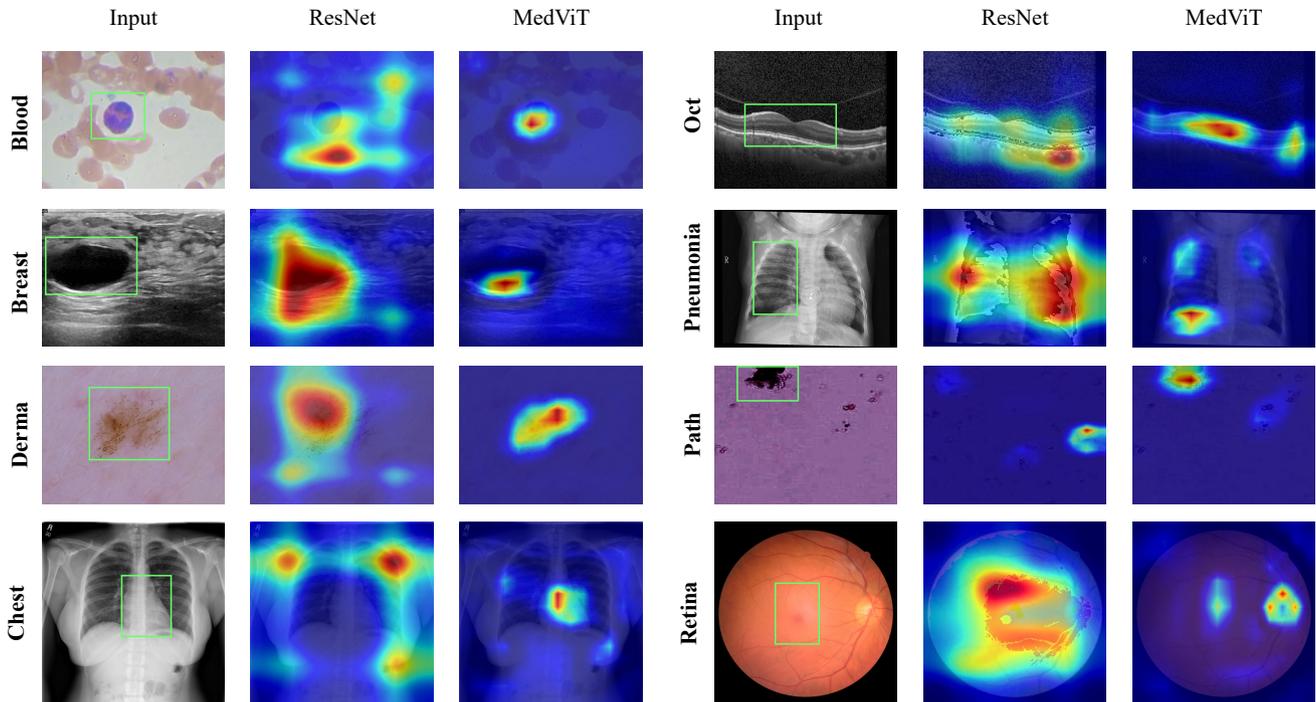


Figure 6: Visual inspection of MedViT-T and ResNet-18 using Grad-CAM on MedMNIST-2D datasets. The green rectangles are used to show a specific part of the image that contains information relevant to the diagnosis or analysis of a medical condition, where the superiority of our proposed method can be clearly seen.

coronal, and sagittal directions within the bounding boxes of the 3D images. The view is the only difference between OrganMNIST and LiST. The images are resized into $1 \times 28 \times 28$ to function as targets for multiclass classification of 11 organs of the body. The training and validation sets are comprised of 115 and 16 CT scans from the source training set, respectively. The 70 CT scans from the source test set are also considered the test set.

4.2. Implementation Details

Our experiment on medical image classification is conducted on the MedMNIST dataset, which is composed of 12 standardized datasets from comprehensively medical resources covering a range of primary data modalities representative of medical images. To make a fair and objective judgment, we follow the same training settings of the MedMNISTv2 [63] without making any changes from the original settings. Specifically, we train all of the MedViT variants for 100 epochs on NVIDIA 2080Ti GPUs, and use a batch size of 128. The images are first resized to a size of 224×224 pixels. We employ an AdamW optimizer [73] with an initial learning rate of 0.001, the learning rate is decayed by a factor set of 0.1 in 50 and 75 epochs. Moreover, we introduce MedViT models at three different network sizes MedViT-T, MedViT-S, and MedViT-L, as shown in Table 1. All of them adopt the best settings investigated in section 2 and are trained for each dataset separately. For MedViT*, we add augmentation in the training phase. The PMC uses the mixture feature normalization with data augmentation at training time for each input image patch.

4.3. Evaluation metric

We report Accuracy (ACC) and Area under the ROC Curve (AUC) as the standard evaluation metrics. In contrast to AUC, which is a free threshold metric used to evaluate continuous prediction scores, ACC uses a threshold-based metric to evaluate discrete prediction labels. Therefore, ACC is more sensitive to class discrepancy than AUC. Because our experiments have many datasets of different sizes and data variety, both ACC and AUC could serve as comprehensive metrics. Although there are many other metrics, to establish a fair comparison, we select ACC and AUC for the benchmarking methods reported in the original publications [63, 74]. We report the results of ACC and AUC for each dataset in table 3. Similar to [63], we average the results over MedMNIST2D and report the average AUC and ACC scores in table 5.

5. Evaluation Results

5.1. Results on Each Dataset

The comparison of the proposal method with previous state-of-the-art (SOTA) methods in terms of the AUC and ACC on each dataset of MedMNIST-2D is shown in Table 3. MedViT outperforms previous SOTA methods by a large margin. Compared to AutoML methods, our MedViT-S shows superior learning ability on both evaluation metrics, observing an increase of 2.3% (AUC) and 3.0% (ACC) in RetinaMNIST and an increase of 1.1% (AUC) and 2.8% (ACC) in TissueMNIST compared to Google AutoML Vision and AutoKeras, respectively. Concretely,

Table 3

Comparison results of the proposed method on the MedMNIST2D in metrics of AUC and ACC. White background shows CNN-based and AutoML methods, while the proposed MedViT are colored in blue. Also **blue** indicates the best result, and **red** displays the second-best.

Methods	PathMNIST		ChestMNIST		DermaMNIST		OCTMNIST		PneumoniaMNIST		RetinaMNIST	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18 (28) [20]	0.983	0.907	0.768	0.947	0.917	0.735	0.943	0.743	0.944	0.854	0.717	0.524
ResNet-18 (224) [20]	0.989	0.909	0.773	0.947	0.920	0.754	0.958	0.763	0.956	0.864	0.710	0.493
ResNet-50 (28) [20]	0.990	0.911	0.769	0.947	0.913	0.735	0.952	0.762	0.948	0.854	0.726	0.528
ResNet-50 (224) [20]	0.989	0.892	0.773	0.948	0.912	0.731	0.958	0.776	0.962	0.884	0.716	0.511
auto-sklearn [75]	0.934	0.716	0.649	0.779	0.902	0.719	0.887	0.601	0.942	0.855	0.690	0.515
AutoKeras [76]	0.959	0.834	0.742	0.937	0.915	0.749	0.955	0.763	0.947	0.878	0.719	0.503
Google AutoML [77]	0.944	0.728	0.778	0.948	0.914	0.768	0.963	0.771	0.991	0.946	0.750	0.531
MedViT-T (224)	0.994	0.938	0.786	0.956	0.914	0.768	0.961	0.767	0.993	0.949	0.752	0.534
MedViT-S (224)	0.993	0.942	0.791	0.954	0.937	0.780	0.960	0.782	0.995	0.961	0.773	0.561
MedViT-L (224)	0.984	0.933	0.805	0.959	0.920	0.773	0.945	0.761	0.991	0.921	0.754	0.552

Methods	BreastMNIST		BloodMNIST		TissueMNIST		OrganAMNIST		OrganCMNIST		OrganSMNIST	
	AUC	ACC										
ResNet-18 (28) [20]	0.901	0.863	0.998	0.958	0.930	0.676	0.997	0.935	0.992	0.900	0.972	0.782
ResNet-18 (224) [20]	0.891	0.833	0.998	0.963	0.933	0.681	0.998	0.951	0.994	0.920	0.974	0.778
ResNet-50 (28) [20]	0.857	0.812	0.997	0.956	0.931	0.680	0.997	0.935	0.992	0.905	0.972	0.770
ResNet-50 (224) [20]	0.866	0.842	0.997	0.950	0.932	0.680	0.998	0.947	0.993	0.911	0.975	0.785
auto-sklearn [75]	0.836	0.803	0.984	0.878	0.828	0.532	0.963	0.762	0.976	0.829	0.945	0.672
AutoKeras [76]	0.871	0.831	0.998	0.961	0.941	0.703	0.994	0.905	0.990	0.879	0.974	0.813
Google AutoML [77]	0.919	0.861	0.998	0.966	0.924	0.673	0.990	0.886	0.988	0.877	0.964	0.749
MedViT-T (224)	0.934	0.896	0.996	0.950	0.943	0.703	0.995	0.931	0.991	0.901	0.972	0.789
MedViT-S (224)	0.938	0.897	0.997	0.951	0.952	0.731	0.996	0.928	0.993	0.916	0.987	0.805
MedViT-L (224)	0.929	0.883	0.996	0.954	0.935	0.699	0.997	0.943	0.994	0.922	0.973	0.806

MedViT steadily improve the performance of visual classification tasks in the MedMNIST-2D benchmark, particularly for PathMNIST, ChestMNIST, DermaMNIST, PneumoniaMNIST and BreastMNIST.

Although a specific architecture designed for a special image format is more accurate in one area, we have designed MedViT by combining efficient blocks to extract local and global features for generalized medical image classification. Also, MedMNIST-2D contains different types of images, including CT, ultrasound, X-ray, and OCT, which are colour or grayscale with different content in the medical domain. Results in Table 3 show that our MedViT performs the classification of medical images well for MedMNIST datasets. Besides, the efficiency in terms of the number of parameters is indicated in Table 5, which will be discussed in the following sections. These results demonstrate that the proposed MedViTs design has effectiveness and good generalization ability.

5.2. Comparison with State-of-the-art Models

We compare our MedViT with the latest state-of-the-art methods (e.g. ViTs, CNNs and hybrid networks) with similar model sizes in Table 4. We achieve a favorable trade-off between complexity and accuracy. Specifically, our MedViT-T achieves a 70.3% Top-1 accuracy compared with CNN models, which is better than EfficientNet-B3 and ResNet-18 with more parameters. Similarly, MedViT-S achieves 73.1% Top-1 accuracy, 5.1% higher than ResNet-50, 2.6% higher than EfficientNet-B4, and 0.5% higher than

ConvNext-T, which are famous CNNs. Moreover, MedViT-L outperforms the ConvNext-B by 0.8%, which has approximately two times more parameters than ours. Furthermore, compared to the pure ViTs, MedViT-T also outperforms PVT-T by a large margin of 6.9%, while the model complexity is much lower. MedViT-S surpasses Twins-SVT-S by 1% with similar number parameters. Finally, compared with recent hybrid methods, MedViT-T beats RVT-Ti by 0.7%. Compared to CvT-13, MedViT-S improve performance by 1.5% while the complexity is similar. MedViT-L also obtains a 0.6% performance gain over RVT-B while enjoying less computation complexity. Experimental results illustrate that the proposed MedViT can effectively handle the classification task.

5.3. Average performance

We compare our method with the average AUC and average ACC over all datasets reported in Table 5. Our models shows average AUCs 93.6%, 94.2%, 93.5% and average ACCs 84%, 85.1%, 84.2% in the total of twelve different datasets obtained by MedViT-T, MedViT-S and MedViT-L, respectively. MedViT-S outperforms all the baseline ResNets and AutoML methods in both average AUC and average ACC by a large margin, demonstrating the advantage of using transformer vision to classify medical image.

In Table 5, we also compare the numbers of parameters of our proposed method with baseline ResNets. Our MedViT shows great superiority in terms of performance

Table 4
Classification performance compared with MedViTs and recent state-of-the-art methods on TissueMNIST.

Network	Image Size	Param (M)	FLOPs (G)	Top-1 (%)
ResNet-18 [20]	224	11.7	1.8	68.1
EfficientNet-B3 [78]	300	12.0	1.8	69.0
DeiT-Ti [79]	224	5.7	1.3	59.5
PiT-Ti [29]	224	4.9	0.7	62.1
PVT-T [8]	224	13.2	1.9	63.4
RVT-Ti [56]	224	8.6	1.3	69.6
MedViT-T	224	10.8	1.3	70.3
ResNet-50 [20]	224	25.6	4.1	68.0
EfficientNet-B4 [78]	380	19.3	4.2	70.5
ConvNeXt-T [23]	224	29.0	4.5	72.6
DeiT-S [79]	224	22.0	4.6	67.0
Swin-T [28]	224	29.0	4.5	71.7
PiT-S [29]	224	23.5	2.9	66.9
PVT-S [8]	224	25.4	4.0	66.7
Twins-SVT-S [80]	224	24.0	2.9	72.1
PoolFormer-S36 [81]	224	31.2	5.0	71.8
CoaT Tiny [13]	224	5.5	4.4	69.3
CvT-13 [12]	224	20.1	4.5	71.6
RVT-S [56]	224	22.1	4.7	71.2
MedViT-S	224	23.6	4.9	73.1
ResNet-152 [20]	224	60.2	11.3	67.5
ConvNeXt-B [23]	224	88.0	15.4	69.1
DeiT-B [79]	224	87.0	17.5	66.9
Swin-B [28]	224	87.8	15.4	68.5
PiT-B [29]	224	73.8	12.5	68.1
PVT-L [8]	224	61.4	9.8	66.8
Twins-SVT-B [80]	224	56.0	8.6	68.7
PoolFormer-M36 [81]	224	56.1	8.8	67.6
CoaT Small [13]	224	22.0	12.6	66.5
CvT-21 [12]	224	32.0	7.1	67.8
RVT-B [56]	224	86.2	17.7	69.3
MedViT-L	224	45.8	13.4	69.9

Table 5
Average performance comparison in standard metrics of average ACC and average AUC over all MedMNIST-2D.

Methods	Params (M)	Avg.	
		AUC	ACC
ResNet-18 (28) [20]	11.2	0.922	0.819
ResNet-18 (224) [20]	11.2	0.925	0.821
ResNet-50 (28) [20]	23.5	0.920	0.816
ResNet-50 (224) [20]	23.5	0.923	0.821
auto-sklearn [75]	-	0.878	0.722
AutoKeras [76]	-	0.917	0.813
Google AutoML [77]	-	0.927	0.809
MedViT-T (224)	10.2	0.936	0.840
MedViT-S (224)	23	0.942	0.851
MedViT-L (224)	45	0.935	0.842

while the model complexity of ours is on par with baseline ResNets.

5.4. Visual inspection of MedViT

To further verify the property of our MedViT, we apply Grad-CAM [82] on the ESA's output in the last ECB to qualitatively inspect MedViT. We visualize the heat maps of the output features from ResNet-18 and MedViT-T in Figure 6. Compared with the baseline ResNet-18, our MedViT covers the relevant locations in the images more precisely and attends less to the background. Moreover, MedViT can better handle the scale variance issue as shown in Derma, Oct and Path. That is, it covers target area accurately

whether they are small, medium, or large in size. In the Retina dataset, it can be seen that our model in the heat map of retinal fundus image can well recognize the direction and area of the specific lesion. Our model well-localized a focal infected area by bacterial infection in the heat map of Chest dataset, while it was also able to delineate the multi-focal lesions in periphery of both upper lungs in the heat map of Pneumonia dataset, which is typical findings for pneumonia.

Such observations demonstrate that introducing the intrinsic IBs of locality and scale-invariance from convolutions to transformers helps MedViT learn capable of simultaneously capturing high-quality and multi-frequency signals. Compared to the conventional approach, our method mitigates the background bias significantly.

5.5. Augmentation and Robustness Evaluation

To evaluate our model against the adversarial attack benchmarks, we adopt a common gradient-based attack method FGSM [83] and a powerful multi-step attack PGD [41] with a step size of $4/255 = 0.015$ and steps $n_{iter} = 5$. For both attackers, the magnitude of the adversarial noise is ϵ of $8/255 = 0.031$. Results in Table 6 demonstrate that different blocks of MedViT architecture have a strong correlation with the adversarial robustness. The proposed MedViT-T and MedViT-T* represent high adversarial robustness under both attack benchmarks. This is ascribed to the Efficient Convolution Block and Patch Momentum Changer, which aims to improve the robustness of medical diagnostic. In ECB, adding depth-wise convolution into feed-forward networks help model to better capture local dependencies within tokens. Moreover, the PMC module utilizes implicit data augmentation at token-level, which forces ViTs to pay special attention towards local features at different stages. We show empirically that MedViT by using these blocks is consistently able to improve robustness and classification accuracy across medical datasets.

The proposed MedViT-T* model achieves superior performance on both admired PGD and FGSM attacks in compared ResNet and baseline MedViT. In detail, MedViT-T* considerably outperforms the counterparts in TissueMNIST with gains of 38.4%, 6.1% on FGSM attack and gains of 30.2%, 7.5% on PGD attack compared to ResNet-18 and MedViT-T, respectively. MedViT-T* also achieves outstanding standard performance (ACC) consistently on four MNIST dataset. Specifically, the Transformer Augmentation Block module brings significant improvements (1.5%, 3.1%, 1.1% and 0.7%) on the Oct, Tissue, Retina and Path MNIST datasets, respectively. This advance is further expanded by our Locally-FeedForward and PMC augmentation. Nonetheless, our MedViT-T* model generally yields the best accuracy/robustness tradeoff.

5.6. Ablation Study

We conduct various ablation experiments to investigate the effectiveness of the critical blocks of our architecture. Firstly, we study the impact of Efficient Convolution Block on robust and clean accuracy in comparison with the most

Table 6

The performance of MedViT-T, MedViT-T* and ResNet-18 on four MedMNIST-2D and two robustness benchmarks. Except for MedViT-T* architecture, we do not make use of any specialized modules or additional fine-tuning procedures.

Methods	OctMNIST			TissueMNIST			RetinaMNIST			PathMNIST		
	ACC	FGSM	PGD									
ResNet-18 (224)	0.763	0.238	0.201	0.681	0.096	0.090	0.493	0.167	0.117	0.909	0.406	0.108
MedViT-T (224)	0.767	0.272	0.249	0.703	0.419	0.317	0.534	0.168	0.145	0.938	0.562	0.224
MedViT-T* (224)	0.782	0.304	0.297	0.734	0.480	0.392	0.545	0.197	0.180	0.945	0.585	0.245

Table 7

Impact of Efficient Convolution Block. Performance of clean accuracy and adversarial robustness under FGSM attack on TissueMNIST.

Block type	Model Complexity		Clean Acc(%)	Robust Acc(%)
	Params (M)	Flops (G)		
Residual Block [20]	10.9	1.3	68.1	22.3
ConvNeXt Block [23]	11.2	1.4	69.7	37.5
PoolFormer Block [81]	10.7	1.1	68.9	29.1
LSA Block [80]	12.7	2.1	69.2	31.8
ECB (ours)	10.8	1.3	70.3	41.9

well-known components. Afterwards, we individually evaluate Patch Moment Changer block at different stages of our architecture. It is important to note that all of our ablation experiments are based on the MedViT-T model on TissueMNIST.

Impact of Efficient Convolution Block. To analyze the effectiveness of our ECB for improving robustness/accuracy of Transformers, we substitute ECB in MedViT with famous blocks, including ConvNext [23] block, Residual block in ResNet [20], PoolFormer Block [81], and LSA block in Twins [80]. We constantly keep other components of our architecture unchanged to build different models under similar complexity. As illustrated in Table 7, our architecture with ECB block achieves the best robustness/accuracy in comparison with prior blocks. In particular, ECB outperforms the ConvNext block (runner-up) by 0.6% in clean and 4.4% in robust accuracy with lower model complexity.

Effect of PMC in Different Stages. To find the best place for Patch Moment Changer in our model, we combine the efficient blocks with PMC in different stages. PMC is applied to 3 different stages of MedViT-T on TissueMNIST to find the best setup. As illustrated in Table 8, PMC works best when applied after the first stage of the 4-stage MedViT-T. We assume PMC helps transformers to capture local information and has a significant advantage at the early stages. In contrast, the last stages already contain a lot of information, which is less impacted by the effect of PMC. In this paper, we adopt PMC augmentation in the first stage.

6. Conclusion

In this paper, we introduce a family of MedViT, a novel hybrid CNN-transformer architecture for medical image classification. Specifically, we combine the local

Table 8

Effect of PMC in Different Stages. Performance (%) of clean accuracy and adversarial robustness under FGSM attack on TissueMNIST. Place of PMC is indicated by ✓ in different stages.

Augmentations			Acc	Rob. Acc
Stage 1	Stage 2	Stage 3		
✗	✗	✗	70.3	41.9
✓	✗	✗	73.4	48.0
✗	✓	✗	72.9	45.4
✗	✗	✓	71.5	44.1

representations and the global features by using robust components. Furthermore, we have devised a novel patch moment changer augmentation that adds rich diversity and affinity to training data. Experiments show that our MedViT achieves state-of-the-art accuracy and robustness on the standard large-scale collection of 2D biomedical datasets. We hope that our model can encourage more researchers and provide inspire for future works on realistic medical deployment.

CRedit authorship contribution statement

Omid Nejati Manzari: Conceptualization, Software, Writing- Original draft preparation, Validation, Resources. **Hamid Ahmadabadi:** Methodology, Writing- Reviewing and Editing. **Hossein Kashiani:** Writing- Reviewing and Editing, Modification for the final layout. **Shahriar B. Shokouhi:** Supervision, Review & Editing. **Ahmad Aya-tollahi:** Supervision, Review & Editing.

References

- [1] C.-M. Lo and P.-H. Hung, "Computer-aided diagnosis of ischemic stroke using multi-dimensional image features in carotid color doppler," *Computers in Biology and Medicine*, vol. 147, p. 105779, 2022.
- [2] W. Hu, C. Li, X. Li, M. M. Rahaman, J. Ma, Y. Zhang, H. Chen, W. Liu, C. Sun, Y. Yao *et al.*, "Gashissdb: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer," *Computers in biology and medicine*, vol. 142, p. 105207, 2022.
- [3] Q. Hu, C. Chen, S. Kang, Z. Sun, Y. Wang, M. Xiang, H. Guan, L. Xia, and S. Wang, "Application of computer-aided detection (cad) software to automatically detect nodules under sdct and ldct scans with different parameters," *Computers in Biology and Medicine*, vol. 146, p. 105538, 2022.

- [4] X. Yang and M. Stamp, "Computer-aided diagnosis of low grade endometrial stromal sarcoma (Igees)," *Computers in Biology and Medicine*, vol. 138, p. 104874, 2021.
- [5] S. Igarashi, Y. Sasaki, T. Mikami, H. Sakuraba, and S. Fukuda, "Anatomical classification of upper gastrointestinal organs under various image capture conditions using alexnet," *Computers in Biology and Medicine*, vol. 124, p. 103950, 2020.
- [6] R. Togo, H. Watanabe, T. Ogawa, and M. Haseyama, "Deep convolutional neural network-based anomaly detection for organ classification in gastric x-ray examination," *Computers in biology and medicine*, vol. 123, p. 103903, 2020.
- [7] 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 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [8] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.
- [9] 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 transactions on pattern analysis and machine intelligence*, 2022.
- [10] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [11] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.
- [12] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.
- [13] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9981–9990.
- [14] J. Guo, K. Han, H. Wu, C. Xu, Y. Tang, C. Xu, and Y. Wang, "Cmt: Convolutional neural networks meet vision transformers," *arXiv preprint arXiv:2107.06263*, 2021.
- [15] L. Ma and L. Liang, "A regularization method to improve adversarial robustness of neural networks for ecg signal classification," *Computers in Biology and Medicine*, vol. 144, p. 105345, 2022.
- [16] B. Li, F. Wu, S.-N. Lim, S. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 383–12 392.
- [17] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5275–5285.
- [18] C. Cao, F. Zhou, Y. Dai, and J. Wang, "A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability," *arXiv preprint arXiv:2212.10888*, 2022.
- [19] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," pp. 11 976–11 986, 2022.
- [24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.
- [25] 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.
- [26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.
- [27] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and 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.
- [28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [29] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 936–11 945.
- [30] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 519–16 529.
- [31] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 259–12 269.
- [32] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.
- [33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [34] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," *arXiv preprint arXiv:2108.05895*, 2021.
- [35] O. N. Manzari, H. Kashiani, H. A. Dehkordi, and S. B. Shokouhi, "Robust transformer with locality inductive bias and feature normalization," *Engineering Science and Technology, an International Journal*, vol. 38, p. 101320, 2023.
- [36] B. Wu, J. Chen, D. Cai, X. He, and Q. Gu, "Do wider neural networks really help adversarial robustness?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 7054–7067, 2021.
- [37] E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "A simple way to make neural networks robust against diverse image corruptions," in *European Conference on Computer Vision*. Springer, 2020, pp. 53–69.
- [38] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.
- [39] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When nas meets robustness: In search of robust architectures against adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 631–640.
- [40] M. Dong, Y. Li, Y. Wang, and C. Xu, "Adversarially robust neural architectures," *arXiv preprint arXiv:2009.00902*, 2020.
- [41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
- [42] Y. Li, Q. Yu, M. Tan, J. Mei, P. Tang, W. Shen, A. Yuille, and C. Xie, "Shape-texture debiased neural network training," *arXiv preprint*

arXiv:2010.05981, 2020.

- [43] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.
- [44] S. Ye, K. Xu, S. Liu, H. Cheng, J.-H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang, and X. Lin, “Adversarial robustness vs. model compression, or both?” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 111–120.
- [45] J. Lin, C. Gan, and S. Han, “Defensive quantization: When efficiency meets robustness,” *arXiv preprint arXiv:1904.08444*, 2019.
- [46] C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le, “Smooth adversarial training,” *arXiv preprint arXiv:2006.14536*, 2020.
- [47] R. Zhang, “Making convolutional networks shift-invariant again,” in *International conference on machine learning*. PMLR, 2019, pp. 7324–7334.
- [48] C. Vasconcelos, H. Larochelle, V. Dumoulin, N. L. Roux, and R. Goroshin, “An effective anti-aliasing approach for residual networks,” *arXiv preprint arXiv:2011.10675*, 2020.
- [49] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, “On the adversarial robustness of vision transformers,” *arXiv preprint arXiv:2103.15670*, 2021.
- [50] K. Mahmood, R. Mahmood, and M. Van Dijk, “On the robustness of vision transformers to adversarial examples,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7838–7847.
- [51] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, “Understanding robustness of transformers for image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 231–10 241.
- [52] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Kastramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta *et al.*, “Adversarial attack vulnerability of medical image analysis systems: Unexplored factors,” *Medical Image Analysis*, vol. 73, p. 102141, 2021.
- [53] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, “Towards evaluating the robustness of deep diagnostic models by adversarial attack,” *Medical Image Analysis*, vol. 69, p. 101977, 2021.
- [54] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *arXiv preprint arXiv:2201.09873*, 2022.
- [55] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and 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.
- [56] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, “Towards robust vision transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 042–12 051.
- [57] N. Park and S. Kim, “How do vision transformers work?” *arXiv preprint arXiv:2202.06709*, 2022.
- [58] L. Kauffmann, S. Ramanoël, and C. Peyrin, “The neural bases of spatial frequency processing during scene perception,” *Frontiers in integrative neuroscience*, vol. 8, p. 37, 2014.
- [59] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pvtv2: Improved baselines with pyramid vision transformer,” *arXiv preprint arXiv:2106.13797*, 2021.
- [60] G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian, “Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 458–467.
- [61] L. Carratino, M. Cissé, R. Jenatton, and J.-P. Vert, “On mixup regularization,” *arXiv preprint arXiv:2006.06049*, 2020.
- [62] S. Chen, E. Dobriban, and J. H. Lee, “A group-theoretic framework for data augmentation,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 9885–9955, 2020.
- [63] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, “Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification,” *arXiv preprint arXiv:2110.14795*, 2021.
- [64] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study,” *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.
- [65] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.
- [66] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
- [67] D. Dataset, “The 2nd diabetic retinopathy-grading and image quality estimation challenge,” 2020.
- [68] K. Qi and H. Yang, “Elastic net nonparallel hyperplane support vector machine and its geometrical rationality,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [69] K. Chen, Y. Mao, H. Lu, C. Zeng, R. Wang, and W.-S. Zheng, “Alleviating data imbalance issue with perturbed input during inference,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 407–417.
- [70] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, “Annotated high-throughput microscopy image sets for validation,” *Nature methods*, vol. 9, no. 7, pp. 637–637, 2012.
- [71] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, “A dataset of microscopic peripheral blood cell images for development of automatic recognition systems,” *Data in brief*, vol. 30, 2020.
- [72] W. Al-Dhabyani, M. Goma, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in brief*, vol. 28, p. 104863, 2020.
- [73] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [74] J. Yang, R. Shi, and B. Ni, “Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 191–195.
- [75] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, “Efficient and robust automated machine learning,” *Advances in neural information processing systems*, vol. 28, 2015.
- [76] H. Jin, Q. Song, and X. Hu, “Auto-keras: An efficient neural architecture search system,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1946–1956.
- [77] E. Bisong, “Google automl: cloud vision,” in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 2019, pp. 581–598.
- [78] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning*, 2019, pp. 6105–6114.
- [79] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*, 2021, pp. 10 347–10 357.
- [80] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, “Twins: Revisiting the design of spatial attention in vision transformers,” *arXiv preprint arXiv:2104.13840*, 2021.
- [81] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, “Metaformer is actually what you need for vision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 819–10 829.
- [82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

- [83] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.