Class-consistent Contrastive Learning Driven Cross-dimensional Transformer for 3D Medical Image Classification

Qikui Zhu[1] , **Chuan Fu**[2] , **Shuo Li**[1]
[1]Department of Biomedical Engineering and
Department of Computer and Data Science, Case Western Reserve University, OH, USA
[2]Department of Computer Science, Chongqing university, Chongqing, China
QikuiZhu@163.com, fuchuan2024@163.com, slishuo@gmail.com

Abstract

Transformer emerges as an active research topic in medical image analysis. Yet, three substantial challenges limit the effectiveness of both 2D and 3D Transformers in 3D medical image classification: 1) Challenge in capturing spatial structure correlation due to the unreasonable flattening operation; 2) Challenge in burdening the high computational complexity and memory consumption due to the quadratic growth of computational complexity and memory consumption for 3D medical data; 3) Challenge in discriminative representation learning, due to data-sensitivity. To address the above challenges, a novel Cross-dimensional Transformer (CdTransformer) and a creative Class-consistent Contrastive Learning (CcCL) are proposed. Specifically, CdTransformer consists of two novel modules: 1) Cross-dimensional Attention Module (CAM), which breaks the limitation that Transformer cannot reasonably establish spatial structure correlation when meeting 3D medical data, meanwhile, reduces the computational complexity and memory consumption. 2) Inter-dimensional Feed-forward Network (IdFN), which addresses the challenge of traditional feed-forward networks not being able to learn depth dimension information that is unique to 3D medical data. CcCL innovatively takes full advantage of the inter-class and intra-class features from the slice-distorted samples to boost Transformer in learning feature representation. CdTransformer and CcCL are validated on six 3D medical image classification tasks. Extensive experimental results demonstrate that CdTransformer outperforms state-of-the-art CNNs and Transformers on 3D medical image classification, and CcCL enables significantly improving Transformer in discriminative representation learning.

1 Introduction

Transformers [Vaswani *et al.*, 2017] have achieved performance breakthroughs in capturing long-range token dependencies and global feature extraction. However, it meets three challenges (Figure.1(a)) when applied to 3D medical image data: 1) The flattening operation of Transformer destroys the 3D spatial structure of data, making it difficult to capture spatial structure correlations and brings irrelevant information and ambiguity to feature representation learning for 3D medical image data. 2) Huge computational complexity and memory consumption due to the computational complexity and memory consumption of Transformer is quadratic to the spatial size. 3) Challenge in learning the feature representation with limited positive and negative samples, particularly for discriminative representation learning, due to transformer-based models are only effective when large training datasets are available. Although many Transformer methods [Fu *et al.*, 2024; Guo *et al.*, 2021; Chen *et al.*, 2021b; Zhu *et al.*, 2022; Zhu *et al.*, 2023a] have achieved encouraging performance in 3D data analysis, the aforementioned challenges remain

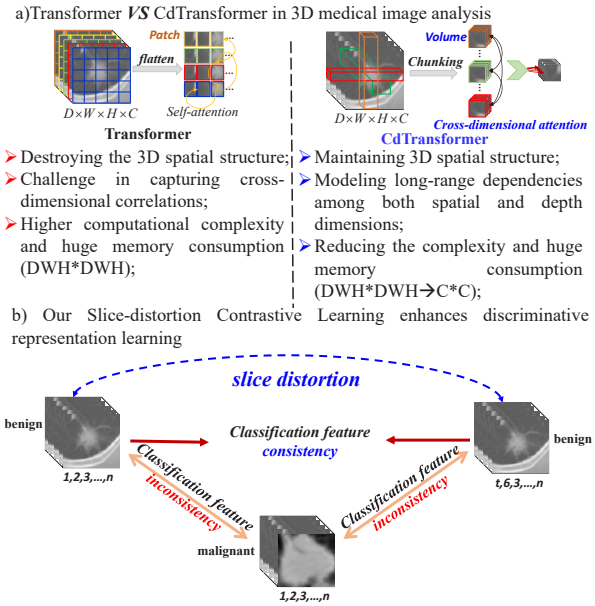

Figure 1: Our CdTransformer and CcCL overcome the disadvantages of Transformers in 3D medical image analysis by maintaining local 3D spatial structure, reducing the complexity and memory consumption, and improving discriminative representation learning.

unresolved and limit the effectiveness of Transformer on 3D medical image analysis.

Existing 3D Transformers are also unable to overcome the aforementioned challenges. For example, the 3D Transformer used in [Zhou *et al.*, 2023] employs Local Volume-based Multi-head Self-attention (LV-MSA) and Global Volume-based Multi-head Self-attention (GV-MSA) to construct feature pyramids for learning representations on 3D volumes. Although LV-MSA reduces computational resources, it can only extract local information. Furthermore, GV-MSA faces the challenge of high computational complexity while learning global feature representations. The Multi-plane and Multi-slice Transformer [Jang and Hwang, 2022] extracts 3D feature representations by constructing attention relationships among multi-plane (axial, coronal, and sagittal) and multi-slice images, but it is unable to avoid the disruption of 3D spatial structure. Additionally, the Multi-plane and Multi-slice Transformer, built upon the Transformer architecture, also encounters the challenge of high computational complexity for 3D input. Despite the various strategies employed by existing methods to integrate Transformers with 3D data, these 3D Transformers still fail to overcome the challenges inherent in 3D medical data. Hence, there is an urgent need for a computationally efficient, 3D structurally aware Transformer in the field of 3D medical image analysis.

We propose a novel Cross-dimensional Transformer (CdTransformer) that consists of 1) a Cross-dimensional Attention Module (CAM) and 2) an Inter-dimensional Feed-forward Network (IdFN) to exploit long-range token dependencies among 3D spatial pixels and extract the global discriminative representations to overcome above challenges and further improve the ability of Transformer on 3D medical image analysis (Figure.1(a)). Specifically, CAM introduces a novel cross-dimensional attention that promotes information sharing and fusion between different dimensions for building connections between 3D spatial pixels, addressing the challenge that Transformer lacking reasonable spatial structure correlation establishes mechanisms. More significantly, CAM converts the quadratic memory consumption of Transformer to linear memory consumption for addressing high complexity and memory consumption problems. IdFN adopts a novel Inter-dimensional attention to activate and fuse the effective features from both spatial and depth views, overcoming the challenge of Feed-forward Network ignoring the depth dimension information.

Apart from architectural novelties, a novel contrastive learning, named Class-consistent Contrastive Learning (CcCL) (Figure.1(b)), is proposed to overcome the limitation that Transformer-based methods are ineffective when meeting limited training datasets [Chen *et al.*, 2020]. Specifically, CcCL innovatively exploits the inter-class feature and intra-class feature from the slice-distorted samples by taking full advantage of Transformer's strengths in long-range dependencies learning. By maximizing the consistent of positive slice-distorted pairs while minimizing the inconsistency of negative slice-distorted pairs, CcCL enables Transformer to learn class-aware discriminative features under limited training datasets. We conduct comprehensive experiments and demonstrate the significance of CcCL in boosting discriminative representation learning with limited positive and negative samples.

Our proposed CcCL and CdTransformer were validated on six 3D medical image classification tasks including five 3D MedMNIST datasets [Yang *et al.*, 2023] and one well-known LIDC-IDRI dataset [Kuan *et al.*, 2017]. Extensive experimental results demonstrate that CdTransformer outperforms state-of-the-art CNNs and Transformers on 3D medical image classification.

Our main contributions include:

- A cross-dimensional Transformer with linear computational complexity and memory consumption has been established for 3D medical image analysis, which addresses three limitations of Transformer in 3D medical image analysis by effectively capturing spatial structure correlations while mitigating the challenges of high computational complexity and memory consumption.
- Our class-consistent contrastive learning innovatively takes full advantage of the inter-class feature and intra-class feature from the slice-distorted samples to boost the effectiveness of the Transformer in learning discriminative representation.
- Our cross-dimensional attention module innovatively establishes the 3D spatial structure correlation in 3D medical image data, where Transformer cannot.
- Our inter-dimensional feed-forward network enables advanced aggregating spatial and depth contexts from both spatial and depth views, which addresses the limitation that the feed-forward Network of Transformer ignores the depth dimension information.
- Experimental results on six 3D medical image datasets demonstrate that our cross-dimensional Transformer serves as a generalized module, exhibiting remarkable capabilities in 3D medical data analysis.

2 Related Work

Many authors try to use Transformer [Dosovitskiy *et al.*, 2020] for 3D medical image analysis [Zhao *et al.*, 2023; Zhu *et al.*, 2023b; Qin *et al.*, 2022]. For example, Xie et al. [Xie *et al.*, 2021] proposed a hybrid model of CNN Transformer, namely CoTr, for 3D medical image segmentation. Inside the model, the deformable Transformer (DeTrans) that employs the deformable self-attention mechanism is introduced to reduce the computational and spatial complexities of modelling the long-range dependency on multi-scale and high-resolution feature maps. Wang et al. [Wang *et al.*, 2022] proposed an fNIRS classification network based on Transformer, named fNIRS-T, for functional near-infrared spectroscopy classification. To explore the spatial-level and channel-level representation of fNIRS signals, two Transformers, fNIRS Spatial-level Transformer (fNIRS-ST) and fNIRS Channel-level Transformer (fNIRS-CT), are employed inside model. fNIRS-ST can extract local brain area features and fNIRS-CT for the hemodynamic response of a single channel. Hatamizadeh et

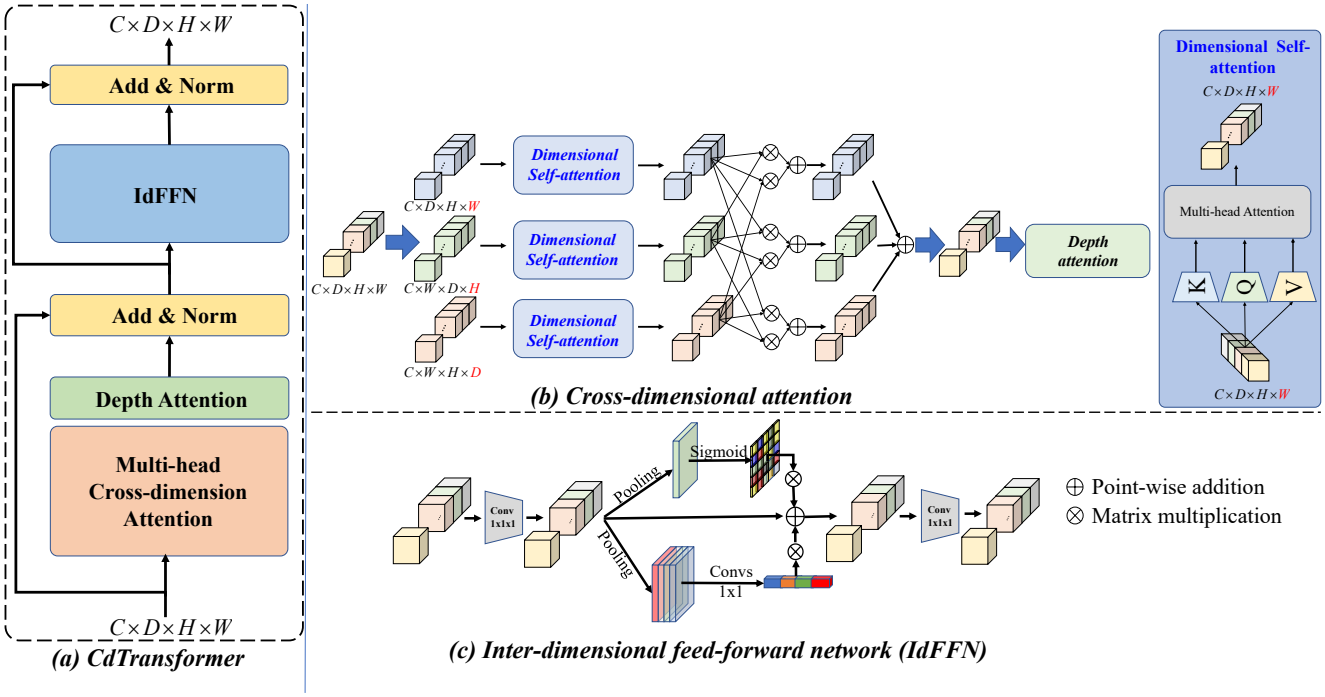

Figure 2: (a) CdTransformer builds cross-dimensional connections through the novel b) cross-dimensional attention module and c) Inter-dimensional Feed-forward Network.

al. [Hatamizadeh *et al.*, 2022] proposed a novel architecture, dubbed as UNEt TRansformers (UNETR), that utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information.

3 Method

3.1 Cross-dimensional Transformer

Cross-dimensional Transformer (CdTransformer) are advanced in 3D medical data analysis through two key innovations: 1) Cross-dimensional Attention Module (CAM); 2) Inter-dimensional Feed-forward Network (IdFN). The key design elements of CAM are cross-dimensional attention and depth attention. The cross-dimensional attention promotes information exchange and fusion between each dimension for building the spatial structure relationship. Depth attention is attached behind CAM which exploits the correlations along depth dimensions. Compared with existing Transformers, CAM has the advantage of capturing spatial structure correlation and converting the quadratic complexity into linear complexity and significantly reduces the computational memory consumption. IdFN is designed for better aggregating and transforming spatial and depth contexts through a novel architecture. IdFN learns local image features from spatially neighboring pixels and exploits the spatial and depth-sensitive features via the attention mechanism, which overcomes the challenge that the feed-forward network of Transformer ignores depth contexts. The two modules complement each other and investigate the long-range dependencies among both spatial and depth dimensions with linear memory consumption.

1) Cross-dimensional Attention Module (CAM)
CAM (Figure.2(b)) can investigate the long-range dependencies among pixel's spatial structure and model the global spatial context with linear memory and computation consumption. Specifically, CAM utilizes three independent cross-dimensional attention to learn global spatial contexts by modeling dimensional correlation rather than the spatial dimension, which enables CAM to build connections among dimensions and absorb the complementary information from spatial structure.

Formally, given a 3D input feature map $\mathrm{X} \in \mathbb{R}^{C\times D\times H\times W}$, where C is the number of channels, D, H, and W represent the spatial dimension. The CdTransformer layer first utilizes three convolution layers to project X into three sequences $[Q;K;V] \in \mathbb{R}^{c\times D\times H\times W}$, where c is the hidden dimension of the input sequences, where Q is the input Query sequence, K and V are the input Key,Value sequence. Afterward, different from the conventional self-attention that computes spatial attention maps (its memory complexity brought by the key-query dot product interaction is quadratic with the spatial resolution of 3D input data), cross-dimensional attention performs attention calculation on three dimension $\{D, H, W\}$ separately by three independent self-attention modules.

$$Y_{x\in\{D,H,W\}} = \mathrm{softmax}(\frac{Q_x K_x^T}{d_x})V_x, \tag{1}$$

To ensures that the contextualized global relationships between pixels are exploited, the pixel-wise aggregation of

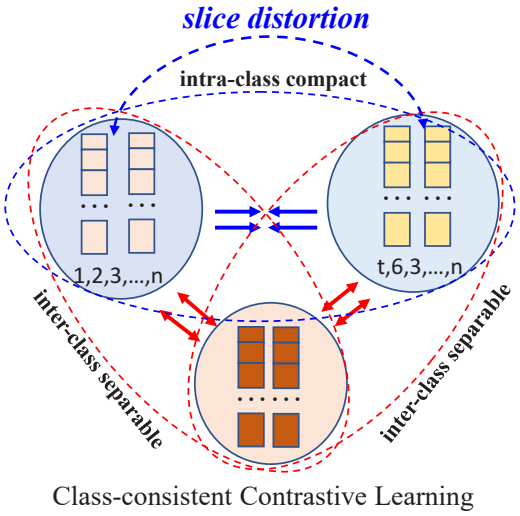

Figure 3: The conception of class-consistent contrastive learning, which pushes inter-class features apart and pushes intra-class features close.

cross-dimensional context is fused via an efficient cross-dimensional fusion layer. Formally,

$$Y'_D = \hat{Y}_D \times \hat{Y}_H + \hat{Y}_D \times \hat{Y}_W, \tag{2}$$

$$Y'_H = \hat{Y}_H \times \hat{Y}_W + \hat{Y}_H \times \hat{Y}_D, \tag{3}$$

$$Y'_W = \hat{Y}_W \times \hat{Y}_D + \hat{Y}_W \times \hat{Y}_H \tag{4}$$

where $Q_D \in R^{\mathrm{D}\times(CHW)}$, $K_D \in R^{\mathrm{D}\times(CHW)}$, $V_D \in R^{\mathrm{D}\times(CHW)}$. d is a learnable scaling parameter to control the magnitude of the dot product. $\hat{Y}_x$ is the reshape of Y_x. And the size of attention map in Y_D, Y_H, Y_W is $\mathbb{R}^{D\times D}$, $\mathbb{R}^{H\times H}$, and $\mathbb{R}^{W\times W}$, respectively.

To overcome the challenge that Transformer only builds the dependencies along spatial dimensions and ignores the relation between depth dimensions, a depth attention module is attached behind the attention module. Formally, three 3D convolution layers with $1 \times 1 \times 1$ kernel size first project $Y_S = [Y'_D \| Y'_H \| Y'_W]$ into three sequences $[Q_C; K_C; V_C] \in R^{DHW\times C}$. Afterward, the depth attention is computed

$$Y_C = \mathrm{softmax}(\frac{Q_C K_C^T}{d_C})V_C \tag{5}$$

Thus, the final output of CAM is calculated as Y_C.

2) Inter-dimensional Feed-forward Network (IdFN)

IdFN (Figure. 2(c)) is designed for boosting spatial and depth contexts aggregating and transforming. Different from the regular feed-forward network (FFN) which consists of two convolution layers, IdFN consists of three parallel streams, one of which is used for learning local image features from spatially neighboring pixels by two stacked convolution layers. The other two streams exploit the spatial and depth-sensitive features via attention mechanism. Formally, given an input feature map $\mathrm{X} \in \mathbb{R}^{C\times D\times H\times W}$, IdFN is formulated as:

$$Y = W_s X + W_d X + X \tag{6}$$

where W_s is spatial-wise attention weight, which is generated by a convolutional operation $\mathrm{w_s} \in \mathbb{R}^{C\times 1\times 1\times 1}$ followed by a sigmoid function.

$$W_s = Sigmoid(w_s X) \tag{7}$$

W_d is depth-wise attention weight, which is generated by a global average pooling operation and two convolutional operations. The global average pooling first performed on the input X to generate the depth-wise statistics $z \in \mathbb{R}^C$. Afterward, a simple gating mechanism with a sigmoid activation is performed on the depth-wise statistics to compute depth-wise attention weights, which is achieved via two convolutional operations and can be formulated as:

$$W_d = Sigmoid(w_2(w_1(z))) \tag{8}$$

where w_1 and w_2 denote two convolutional operations.

3.2 Class-consistent Contrastive Learning

CcCL (Figure. 3) innovatively exploits the inter-class feature and intra-class feature from the slice-distorted samples by taking full advantage of the Transformer's strengths in long-range dependencies, which boosts Transformer in discriminative feature representation learning and addresses the challenge of Transformer being ineffective in limited training datasets. Specifically, given one sample x_i, the corresponding positive sample x'_i can be obtained by randomly selecting one dimension and randomly changing the order of slices as shown in Figure.1(b). As the sample and generated sample $\{x_i, x'_i\}$ belong to one category, the category-aware features should be consistent. CcCL enables pushing inter-category feature apart and pulling intra-category feature close. With the guidance of CcCL, Transformer can learn the category-aware discriminate features. The formulation of CcCL can be:

$$L_{CcCL} = -\sum_{i\in I} \log \frac{\exp(z_{x_i} \times z_{x'_i}/\tau)}{\sum_{y\in A} \exp(z_{x_i} \times z_y/\tau)} \tag{9}$$

where $z_x = Proj(Enc(x))$, the $\times$ symbol denotes the inner product, τ is a scalar temperature parameter, I is the index of an arbitrary sample, A represents sample from different categories. Remarkable, our CcCL enables producing unlimited positive samples, which addresses the challenge of sample choosing in contrastive learning and enhances Transformer in feature representation learning by creatively capturing category-aware correlated features.

4 Experiment

4.1 Datasets and Implementation Details

1) 3D MedMNIST Dataset: Five 3D standardized medical datasets, including AdrenalMNIST3D, NoduleMNIST3D, VesselMNIST3D, SynapseMNIST3D, FractureMNIST3D, with diverse classification tasks (binary classes, multi-classes, and multi-label) of 3D MedMNIST dataset [Yang *et al.*, 2023] are used in this paper. These 3D medical data as the

Methods	Nodule		Fracture		Adrenal		Vessel		Synapse	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18 + 2.5D [He *et al.*, 2016]	0.838	0.835	0.587	0.451	0.718	0.772	0.748	0.846	0.634	0.696
ResNet-18 + 3D [He *et al.*, 2016]	0.863	0.844	0.712	0.508	0.827	0.721	0.874	0.877	0.820	0.745
ResNet-50 + 2.5D [He *et al.*, 2016]	0.835	0.848	0.552	0.397	0.732	0.763	0.751	0.877	0.669	0.735
ResNet-50 + 3D [He *et al.*, 2016]	0.875	0.847	0.725	0.494	0.828	0.745	0.907	0.918	0.851	0.795
ResNet-50 + ACS [Yang *et al.*, 2023]	0.886	0.841	0.750	0.517	0.828	0.758	0.912	0.858	0.719	0.709
auto-sklearn [Feurer *et al.*, 2015]	0.914	0.874	0.628	0.453	0.828	0.802	0.910	0.915	0.631	0.730
AutoKeras [Jin *et al.*, 2019]	0.844	0.834	0.642	0.458	0.804	0.705	0.773	0.894	0.538	0.724
DWT-CV [Cheng *et al.*, 2022]	0.912	0.912	0.723	0.531	0.866	0.812	0.905	0.912	-	-
MDANet [Huang *et al.*, 2022]	0.860	0.868	-	-	0.840	0.815	0.901	0.929	0.712	0.750
ACS [Yang *et al.*, 2023]	0.873	0.847	0.714	0.497	0.839	0.754	0.930	0.928	0.705	0.722
CdTransformer (Ours)	0.919	0.886	0.716	0.517	0.878	0.815	0.919	0.893	0.872	0.818
CdTransformer + CcCL (Ours)	**0.943**	0.903	0.724	0.529	**0.884**	**0.836**	**0.959**	**0.929**	**0.879**	**0.832**

Table 1: Quantitative evaluation results show that CdTransformer achieves the highest AUC and ACC compared with state-of-the-art methods on 3D MedMNIST dataset.

	AUC(%)	ACC(%)	Precision(%)	Sensitivity(%)	F1-Score(%)
Multi-crop CNN [Shen *et al.*, 2017]	93.0	87.14	–	77.0	–
MV-KBC [Samala *et al.*, 2018]	95.70±0.24	91.60±0.15	87.75±0.24	86.52±0.25	87.13±0.16
MSCS-DeepLN [Wu *et al.*, 2018]	94.00±0.25	92.65±0.26	90.39±0.93	85.58±0.94	87.91 ±0.43
MK-SSAC [Xie *et al.*, 2019]	95.81±0.19	92.53±0.05	–	84.94±0.17	–
HSCNN [Shen *et al.*, 2019]	85.6±2.6	84.2±2.5	–	70.5± 4.5	–
Local-Global [Al-Shabi *et al.*, 2019a]	95.62±0.02	88.46±0.04	87.38±0.07	88.66±0.06	88.37±0.04
Gated-Dilated [Al-Shabi *et al.*, 2019b]	95.14±0.03	92.57±0.03	91.85±0.05	92.21±0.04	92.60±0.03
Swarm [de Pinho Pinheiro *et al.*, 2020]	–	93.71	93.53	92.96	–
3D DPN (Ensemble) [Jiang *et al.*, 2020]	–	90.24	–	92.04	90.45
MVCS [Zhu *et al.*, 2022]	91.25	91.25	91.59	89.10	90.19
CdTransformer (Ours)	92.46±0.02	92.53±0.02	93.16±0.05	91.28±0.03	92.04 ±0.02
CdTransformer + CcCL (Ours)	93.82±0.02	**93.98±0.02**	**95.82±0.03**	91.28±0.03	**93.43 ±0.02**

Table 2: Quantitative evaluation results show that CdTransformer achieves the highest Accuracy, Precision, and F1-score compared with state-of-the-art methods on LIDC-IDRI dataset.

benchmark in medical image classification tasks. AdrenalMNIST3D consists of 1584 left and right adrenal glands CT images that are used to distinguish normal adrenal glands from adrenal masses. NoduleMNIST3D consists of 1849 lung nodule chest CT images used to classify two types of malignancy levels. VesselMNIST3D consists of 1909 brain vessels collected by reconstructing MRI images, which are used for binary classification of the aneurysm and healthy vessel segments. The SynapseMNIST3D is a new 3D volume dataset to classify whether a synapse is excitatory or inhibitory.

2) LIDC-IDRI Dataset: The LIDC-IDRI dataset [Kuan *et al.*, 2017] contains 1018 computed tomography (CT) scans from 1010 patients altogether collated from seven academic centers across the United States (US). The slice thicknesses of the CT scans range from 0.45 to 5.0 mm. Each CT scan was annotated by four experienced thoracic radiologists. In this study, we follow previous works [Al-Shabi *et al.*, 2019a; Al-Shabi *et al.*, 2022], the annotated nodules of size smaller than 3mm, slice spacing inconsistent, or missing slices are removed. Finally, there are a total of 837 nodules left, of which 442 nodules are benign and 395 nodules are malignant.

3) Implementation Details: The 3D medical data classification network consists of 4 stages. CdTransformer is used in stage-1 to 4, the number of CdTransformer blocks is [2, 3, 3, 4], the attention heads number in CAM are [2, 4, 8, 16]. The channel expansion factor is 4. Between each stage, a 3D convolution layer with $4 \times 4 \times 4$ kernel size and step size 2 is used for downsampling. Behind the CdTransformer blocks, a classification head, and a projection head are attached.The network is pre-trained under the supervision of L_{CcCL}. For the LIDC-IDRI dataset, all the lung nodules were cropped around the centers of the lung nodules with size $32 \times 32 \times 32$ pixels and normalized by the z-score standardization method (mean value is -400, std value is 750). The randomly adding Gaussian noise, horizontal flip, vertical flip, and z-axis flip are utilized for data augmentation. The framework is implemented on Pytorch and four A4000 GPUs with 16 GB memory, Adam optimizer with a minibatch size of 32 was applied for optimization. The learning rate and weight decay were set to 1e-4 and 0.01, respectively.

4.2 Comparison with State-of-the-art Methods

The quantitative results in Table 1 demonstrate that CdTransformer achieved the best quantitative results on 3D MedM-

	AUC	ACC	Precision	Sensitivity	F1-Score
Transformer [Vaswani *et al.*, 2017]	79.11	78.31	70.59	92.31	80.00
Swin Transformer [Liu *et al.*, 2021]	85.93	85.54	80.00	92.31	85.71
CrossViT [Chen *et al.*, 2021a]	87.06	86.75	81.82	92.31	86.75
ViT [Arnab *et al.*, 2021]	85.87	81.08	10.91	82.19	85.63
SSAN [Guo *et al.*, 2021]	82.63	83.13	87.88	74.36	80.56
MobileViT [Mehta and Rastegari, 2021]	91.32	91.57	**94.44**	87.18	90.67
MMTransformer [Jang and Hwang, 2022]	89.04	89.16	89.47	87.18	88.31
3D Transformer [Zhou *et al.*, 2023]	88.90	89.16	91.67	84.62	88.00
CdTransformer	**94.03**	**93.98**	92.50	**94.87**	**93.76**

Table 3: Quantitative evaluation results of various Transformers on LIDC-IDRI dataset (the 5th fold).

Method	2D/3D	Attention type	Computational complexity	Memory consumption
Transformer	2D	MSA	$2(DHW)^2C + 4DHWC^2$	$(DHW)^2$
ViT [Dosovitskiy *et al.*, 2020]	2D	MSA	$2(DHW)^2C + 4DHWC^2$	$(DHW)^2$
Swin-Transformer [Liu *et al.*, 2021]	2D	MSA (path)	$2(hw)^2DHWC + 4DHWC^2$	$hwDHW$
MMTransformer [Jang and Hwang, 2022]	2D	MSA	$2(HW+HD+WD)^2C$ $+ 4(HW+HD+WD)C^2$	$(HW+HD+DW)^2$
3D Transformer [Zhou *et al.*, 2023]	3D	LV-MSA	$2hwdHWDC + 4DHWC^2$	$hwdDHW$
		GV-MSA	$2(DHW)^2C + 4DHWC^2$	$(DHW)^2$
CdTransformer	3D	Cd-MSA	$6HWCD^2 + 6DWCH^2$ $+ 6DHCW^2$	$D^2 + H^2 + W^2$
		Depth MSA	$6HWDC^2$	C^2

Table 4: The computational complexity and memory consumption of various Transformers. D, H, W represent the volume of 3D medical data. d, h, w is the size of sub-volume.

NIST dataset. Specifically, compared with state-of-the-art methods, our CdTransformer achieved the best AUC, and ACC on four datasets, which demonstrates that CdTransformer has an advantage in capturing the category-aware features through establising long-range pixel dependencies among three dimensions. Meanwhile, as shown in Table 1, with the assistance of CcCL, our CdTransformer further improved the performance, which also proves the superiority of our CcCL in exploiting discriminative features.

The results on LIDC-IDRI database are shown in Table 2. As it can be seen from Table 2, our CdTransformer achieves the highest Accuracy, Precision, and F1-score compared with state-of-the-art methods, which represents our model has distinct advantages over compared methods in 3D nodule representation learning. Meanwhile, those results confirmed that 1) absorbing the global spatial and depth contextual information could further improve the performance; 2) building the dependencies correlations along both spatial and depth dimensions could assist the model to capture global and significant contextual information; 3) CcCL has advantages in improving the performance of CdTransformer.

4.3 Comparison with State-of-the-art Transformer

We also compared CdTransformer with with state-of-the-art Transformers, inclduing Transformer [Vaswani *et al.*, 2017], MobileViT [Mehta and Rastegari, 2021], Swin Transformer [Liu *et al.*, 2021], CrossViT [Chen *et al.*, 2021a], SSAN [Guo *et al.*, 2021], MMTransformer [Jang and Hwang, 2022], and 3D Transformer [Zhou *et al.*, 2023] in 3D medical image classification from the aspects of effectiveness, memory consumption and computational complexity.

Experimental results demonstrate CdTransformer outperforms state-of-art Transformers. From Table. 3, we can notice that the Transformer (ViT) has a poor performance. The major reason is that the vanilla self-attention module cannot effectively and reasonably model long-range dependencies among both spatial and depth dimensions of 3D medical image data. An interesting phenomenon is that SSAN obtained the second worst results. Although SSAN could investigate the relationship between spatial and depth correlations, it lacks the ability to exploit the correlations between dimensions. MobileViT has less computational complexity and achieved the best performance. These results demonstrate the necessity of reducing memory consumption and computational complexity of Transformer in 3D medical image classification and point to three aspects that Transformer needs to face: 1) The less memory consumption and computational complexity are advanced in improving the performance of the Transformer. 2) Effective and reasonable modeling of long-range dependencies among both spatial and depth dimensions are significant to 3D medical image data analysis. 3) It makes perfect sense to boost the effectiveness of the Transformer in learning long-range dependencies and discriminative representation.

CdTransformer significantly reduces computational complexity and memory consumption. As show in Table. 4, given an input data $X \in \mathbb{R}^{C\times D\times H\times W}$, the per-layer complexity of the vanilla self-attention module is $2(DHW)^2C$ +

Baseline	Self-attention		FFN		CcCL	AUC (%)	ACC (%)	Precision (%)	Sensitivity (%)	F1-Score (%)
	SA	CA	FFN	IdFN						
✓	✗	✗	✗	✗	✗	87.06	86.75	81.82	92.31	86.75
✓	✓	✗	✓	✗	✗	91.61	91.57	90.00	92.31	91.14
✓	✗	✓	✓	✗	✗	91.75	91.57	88.10	94.87	91.36
✓	✗	✓	✗	✓	✗	92.45	92.77	**97.14**	87.18	91.89
✓	✗	✓	✗	✓	✓	**94.03**	**93.98**	92.50	**94.87**	**93.76**

Table 5: Quantitative evaluation results of baseline with various configures on LIDC-IDRI dataset.

$4DHWC^2$ and the attention metric size is $DHW \times DHW$. The per-layer complexity of three dimensions inside the CdTransformer is $6HWCD^2$, $6DWCH^2$, $6DHCW^2$ and the attention metric size is $D \times D$, $H \times H$, and $W \times W$, respectively. The complexity of depth dimension is $6DHWC^2$, the attention metric size is $C \times C$. These results demonstrate that CdTransformer reduces computational complexity and memory consumption. What's more, while both 3D Transformer [Zhou *et al.*, 2023] and MMTransformer [Jang and Hwang, 2022] are tailored for 3D medical image datasets, they adopt different approaches for 3D feature extraction. In the case of 3D Transformer, it employs two types of self-attention, leading to heightened computational complexity and increased memory consumption. On the other hand, MMTransformer focuses on extracting 3D structure information by combining insights from three distinct views of the 3D data. However, this fusion process may compromise the inherent 3D structure, and similar to 3D Transformer, the self-attention mechanism in MMTransformer comes with comparable computational demands and memory

4.4 Effect Analysis using CdTransformer

The effectiveness of each part of the CdTransformer is also demonstrated. We analyze the influence of each part on the classification results by ablation studies: *1) 3D ResNet (baseline)*: The classification is achieved by directly using 3D ResNet without attention mechanism as the baseline. *2) Baseline + Transformer*: The Transformer (ViT) is used in each stage of 3D ResNet for capturing long-range pixel dependencies and global feature extracting. *3) Baseline + CAM + FFN*: Different from 2), here the vanilla self-attention module inside Transformer is replaced by a dimension-mutual attention module. Meanwhile, to directly prove that the dimension-mutual attention module has the advancement in investigating the long-range dependencies among pixels and modeling the global spatial context, we use the normal feed-forward network (FFN) for feature transformation. *4) Baseline + CAM + IdFN*: Based on setting 3), we replaced the FFN module with our proposed Inter-dimensional feed-forward network (IdFN). By comparing it with setting 3), we can evaluate the effectiveness of IdFN. *5) CdTransformer*: Our proposed Cross-dimensional Transformer network with CcCL.

Table 5 lists the classification performance of the methods described above. From the Table 5, we can notice that the innovations of CAM and IdFN bring significant enhancements, and the baseline has a poor performance. The major reason is that the typical convolution block construed by stacked convolution layers cannot model long-range dependencies among both spatial and depth dimensions of nodule. The spatial attention module assists the baseline model to obtain a 4.55%, 4.82%, 8.18% and 4.39% improvement on AUC, ACC, Precision, and F1-Score, respectively. When utilizing CAM, the AUC, ACC, Precision, Sensitivuty, and F1-Score achieve 4.69%, 4.82%, 6.28%, 2.56% and 4.61%, respectively. Those improvements proved the effectiveness of the CAM in modeling long-range dependencies among pixels and capturing global and significant contextual information. What's more, when using CAM and IdFN simultaneously, the AUC, ACC, Precision, and F1-Score are increased by 5.39%, 6.02%, 15.32%, and 5.14%, which reveals the advantage of further aggregating and transforming spatial and depth contexts. The above extensive experiments with promising results reveal the power of the CAM and IdFN and its significance in improving the performance of the Transformer.

4.5 Significance of CcCL

The significance of CcCL is proved in Table 1, Table 2 and Table 5. The CcCL assists CdTransformer to obtain an improvement on AUC, ACC, Precision, and F1-Score on the six 3D medical datasets. The CcCL assists CdTransformer in obtaining a 4.55%, 4.82%, 8.18%, and 4.39% improvement on AUC, ACC, Precision, and F1-Score on LIDC-IDRI dataset. What's more, in the five 3D MedMNIST, CcCL also improved the performance of model. All of experimental results demonstrate that CcCL improves the performance of Transformer.

5 Conclusion

Our CdTransformer addressed the limitation of Transformer in 3D medical image classification through a novel cross-dimensional attention module and an inter-dimensional feed-forward network. Cross-dimensional attention module and inter-dimensional feed-forward network modules promote information exchange and fusion between dimensions for modeling the long-range dependencies among three dimensions with linear memory consumption and computational complexity. Additionally, our class-consistent contrastive learning boosts Transformer in learning feature representation. Extensive experimental results on six 3D medical image data demonstrate that CdTransformer outperforms the state-of-the-art CNNs and Transformers in 3D medical image classification. And our class-consistent contrastive learning can significantly improve the performance of Transformer.

References

[Al-Shabi *et al.*, 2019a] Mundher Al-Shabi, Boon Leong Lan, Wai Yee Chan, Kwan-Hoong Ng, and Maxine Tan. Lung nodule classification using deep local–global networks. *International journal of computer assisted radiology and surgery*, 14(10):1815–1819, 2019.

[Al-Shabi *et al.*, 2019b] Mundher Al-Shabi, Hwee Kuan Lee, and Maxine Tan. Gated-dilated networks for lung nodule classification in ct scans. *IEEE Access*, 7:178827–178838, 2019.

[Al-Shabi *et al.*, 2022] Mundher Al-Shabi, Kelvin Shak, and Maxine Tan. Procan: Progressive growing channel attentive non-local network for lung nodule classification. *Pattern Recognition*, 122:108309, 2022.

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021a] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. @inproceedingschen2021crossvit, title=Crossvit: Cross-attention multi-scale vision transformer for image classification, author=Chen, Chun-Fu Richard and Fan, Quanfu and Panda, Rameswar, booktitle=Proceedings of the IEEE/CVF international conference on computer vision, pages=357–366, year=2021 : Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[Chen *et al.*, 2021b] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Cheng *et al.*, 2022] Jianhong Cheng, Hulin Kuang, Qichang Zhao, Yahui Wang, Lei Xu, Jin Liu, and Jianxin Wang. Dwt-cv: Dense weight transfer-based cross validation strategy for model selection in biomedical data analysis. *Future Generation Computer Systems*, 135:20–29, 2022.

[de Pinho Pinheiro *et al.*, 2020] Cesar Affonso de Pinho Pinheiro, Nadia Nedjah, and Luiza de Macedo Mourelle. Detection and classification of pulmonary nodules using deep learning and swarm intelligence. *Multimedia Tools and Applications*, 79(21):15437–15465, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Feurer *et al.*, 2015] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28, 2015.

[Fu *et al.*, 2024] WT Fu, QK Zhu, N Li, YQ Wang, SL Deng, HP Chen, J Shen, LY Meng, and Z Bian. Clinically oriented cbct periapical lesion evaluation via 3d cnn algorithm. *Journal of Dental Research*, 103(1):5–12, 2024.

[Guo *et al.*, 2021] Xudong Guo, Xun Guo, and Yan Lu. Ssan: Separable self-attention network for video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12618–12627, 2021.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2022] Yixiang Huang, Lifu Zhang, and Ruoxi Song. Semantic attention guided multi-dimension information complementary network for medical image classification. 2022.

[Jang and Hwang, 2022] Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20718–20729, 2022.

[Jiang *et al.*, 2020] Hanliang Jiang, Fei Gao, Xingxin Xu, Fei Huang, and Suguo Zhu. Attentive and ensemble 3d dual path networks for pulmonary nodules classification. *Neurocomputing*, 398:422–430, 2020.

[Jin *et al.*, 2019] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1946–1956, 2019.

[Kuan *et al.*, 2017] Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, and Vijay Chandrasekhar. Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge. *arXiv preprint arXiv:1705.09435*, 2017.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[Mehta and Rastegari, 2021] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2021.

[Qin *et al.*, 2022] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022.

[Samala *et al.*, 2018] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Caleb D Richter, and Kenny H Cha. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE transactions on medical imaging*, 38(3):686–696, 2018.

[Shen *et al.*, 2017] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017.

[Shen *et al.*, 2019] Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert systems with applications*, 128:84–95, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2022] Zenghui Wang, Jun Zhang, Xiaochu Zhang, Peng Chen, and Bing Wang. Transformer model for functional near-infrared spectroscopy classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2559–2569, 2022.

[Wu *et al.*, 2018] Hongbo Wu, Chris Bailey, Parham Rasoulinejad, and Shuo Li. Automated comprehensive adolescent idiopathic scoliosis assessment using mvc-net. *Medical image analysis*, 48:1–11, 2018.

[Xie *et al.*, 2019] Yutong Xie, Jianpeng Zhang, and Yong Xia. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest ct. *Medical image analysis*, 57:237–248, 2019.

[Xie *et al.*, 2021] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[Yang *et al.*, 2023] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[Zhao *et al.*, 2023] Shen Zhao, Jinhong Wang, Xinxin Wang, Yikang Wang, Hanying Zheng, Bin Chen, An Zeng, Fuxin Wei, Sadeer Al-Kindi, and Shuo Li. Attractive deep morphology-aware active contour network for vertebral body contour extraction with extensions to heterogeneous and semi-supervised scenarios. *Medical Image Analysis*, 89:102906, 2023.

[Zhou *et al.*, 2023] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[Zhu *et al.*, 2022] Qikui Zhu, Yanqing Wang, Xiangpeng Chu, Xiongwen Yang, and Wenzhao Zhong. Multi-view coupled self-attention network for pulmonary nodules classification. In *Proceedings of the Asian Conference on Computer Vision*, pages 995–1009, 2022.

[Zhu *et al.*, 2023a] Qikui Zhu, Yihui Bi, Danxin Wang, Xiangpeng Chu, Jie Chen, and Yanqing Wang. Coordinated transformer with position\& sample-aware central loss for anatomical landmark detection. *arXiv preprint arXiv:2305.11338*, 2023.

[Zhu *et al.*, 2023b] Qikui Zhu, Lei Yin, Qian Tang, Yanqing Wang, Yanxiang Cheng, and Shuo Li. Dcaug: Domain-aware and content-consistent cross-cycle framework for tumor augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–347. Springer, 2023.