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

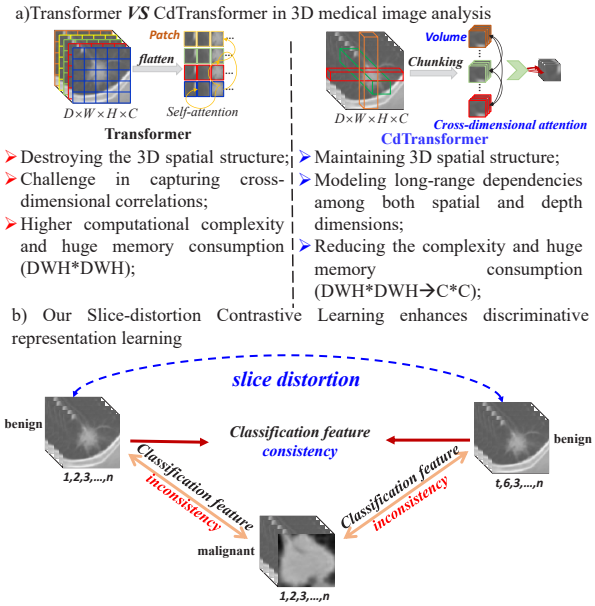

Figure 1: Our CdTransformer and CcCL overcome the disadvantages of Transformers in 3D medical image analysis by maintaining local 3D spatial structure, reducing the complexity and memory consumption, and improving discriminative representation learning.

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

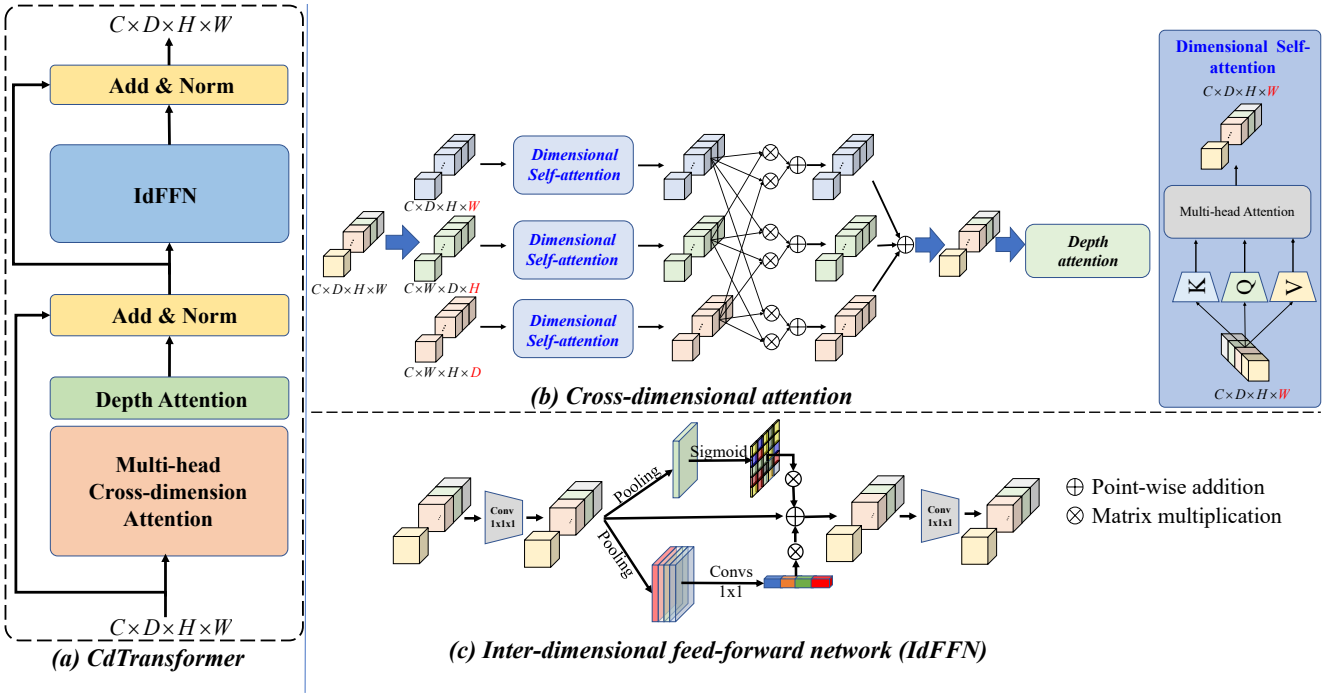

Figure 2: (a) CdTransformer builds cross-dimensional connections through the novel b) cross-dimensional attention module and c) Inter-dimensional Feed-forward Network.

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

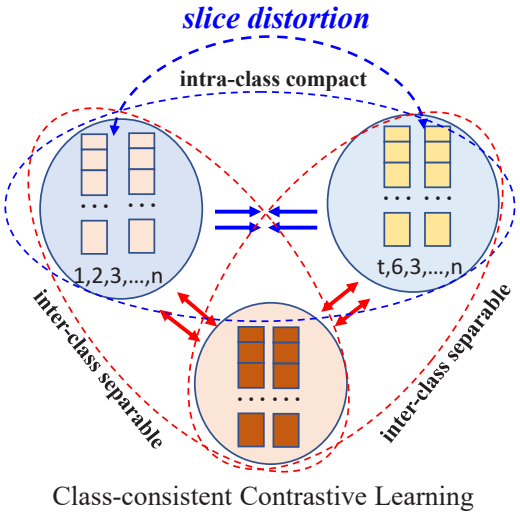

Figure 3: The conception of class-consistent contrastive learning, which pushes inter-class features apart and pushes intra-class features close.

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