Medical Vision Seminar

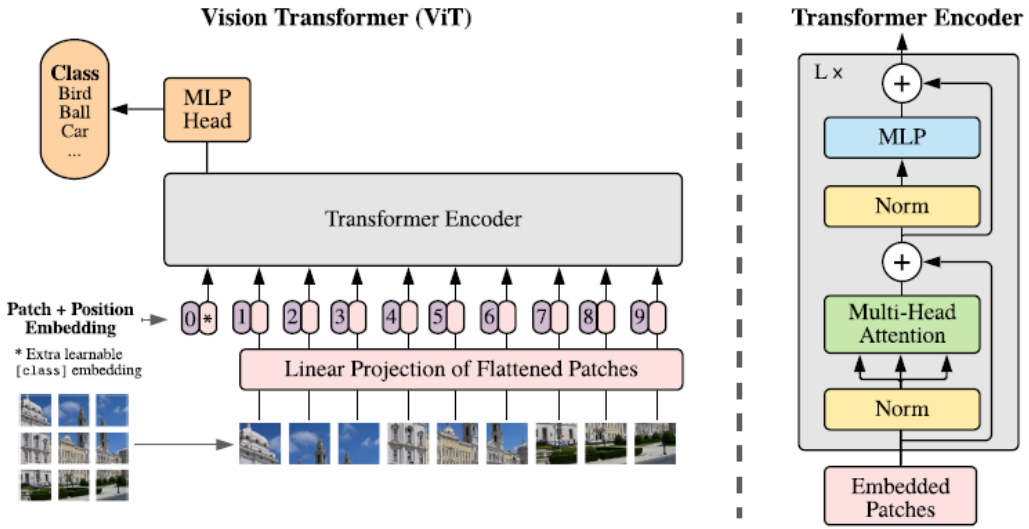
——Wei Lou

（MICCAI 2021）MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus (眼底) Image Classification

—— Shuang Yu, Kai Ma, Qi Bi, Cheng Bian, Munan Ning, Nanjun He, Yuexiang Li, Hanruo Liu, and Yefeng Zheng

1. Motivation

Two shortages of ViT: 1. Large training dataset; 2. Not using patch representation



2. Goal

Design a framework to perform Fundus image classification using limited data and patch representations.

3. Framework

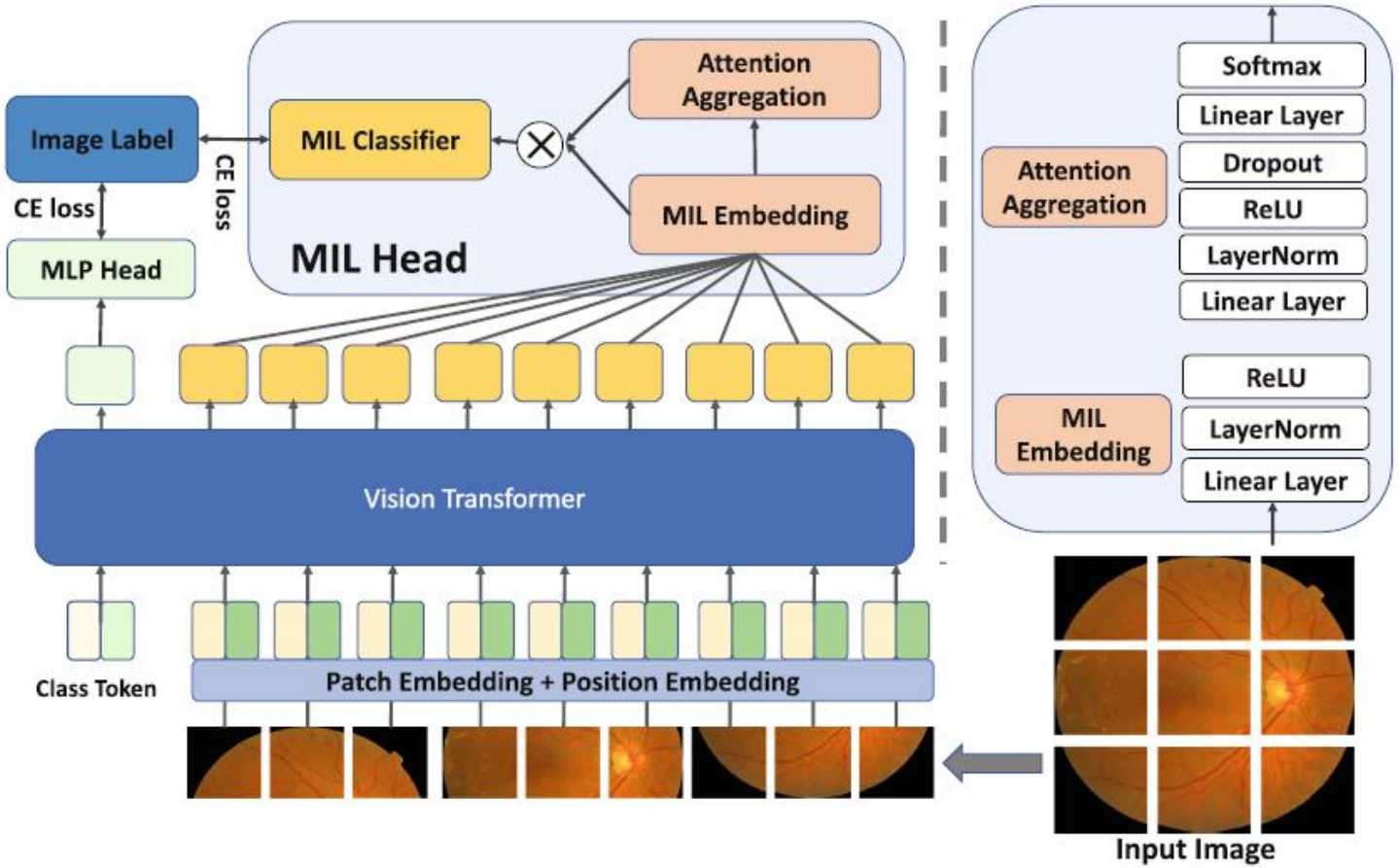


Image Label
CE loss
MLP Head
CE loss
MIL Classifier
Attention Aggregation
MIL Embedding
MIL Head
Vision Transformer
Class Token
Patch Embedding + Position Embedding
Input Image
Attention Aggregation
Softmax
Linear Layer
Dropout
ReLU
LayerNorm
Linear Layer
MIL Embedding
ReLU
LayerNorm
Linear Layer

3.1 Multiple Instance Learning (MIL)

In MIL scheme, it regards an image as a bag, which consists of a set of instances either in the format of pixels or image patches (share the same label). The bag-instance relationship presented in MIL highly resembles that of the image-patch relationship in the Vision Transformer.

Three Steps:

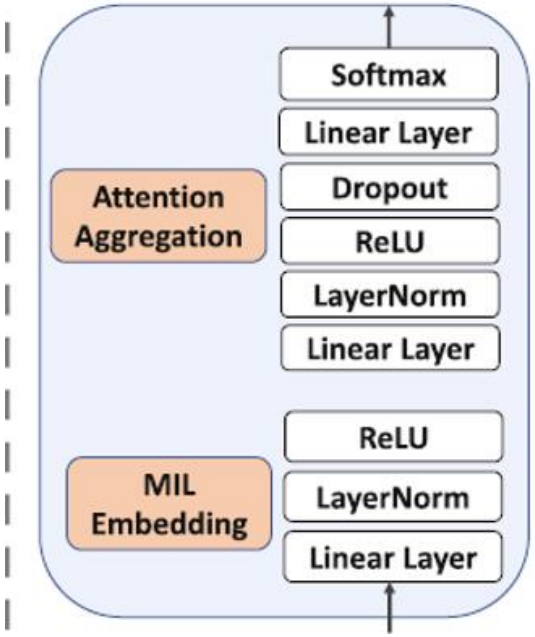
1. Building a lower-dimension embedding for ViT features from individual patches.
2. Use an aggregation function to obtain the bag representation.
3. A bag-level classifier to obtain the final bag-level probabilities.

3.1.1 Lower-dimension embedding

The image can be defined as $X = \{x_1, x_2, \cdots, x_N\}$
And v_i represents the feature vector with dimension of D of instance *i* after transformer encoder. M is the lower dimension.

$$h_i = max(Norm(W^T v_i), 0), \quad h_i \in \mathcal{H}, \mathcal{H} = \mathbb{R}^M,$$

$$Norm(x) = \frac{x - E(x)}{Var(x) + \epsilon} * \gamma + \beta,$$



3.1.2 Attention Aggregation Function

$$\alpha_i = softmax(W_2^T max(Norm(W_1 h_i^T), 0)),$$

An attention module with two linear layers is utilized in our framework to extract the spatial weight matrix for the instance embedding. The attention weights the are assigned to the instance embeddings to highlight the different contribution of different instances. The bag-level representation then is fed into the bag-level classifier.

$$\mathcal{R} = \sum_{i=1}^{N} \alpha_i h_i.$$

3.2 Framework Pre-training and Fine-Tuning

Vision Transformer is trained using a large fundus image classification database (345271).
Five common retinal conditions are labeled, including normal (208,733), diabetic retinopathy (DR, 糖尿病视网膜病变)(38,284), agerelated macular degeneration (老年性黄斑变性) (21,962), glaucoma (青光眼) (24,082), and cataract (白内障) (67,230).

And MIL-VT is then fine-tuned with downstream disease classification tasks. (~5000 images)
APTOS2019 contains a total of 3,662 fundus images for DR grading, with five label categories ranging for 0 to 4, representing the DR severity. The RFMiD2020 dataset contains 1,900 images, with binary disease label being provided: 0 for normal images MIL-VT 51 and 1 for image with pathologies.

Experiments

3.1 Comparison with state-of-the-art methods

Method	AUC	Acc	F1	Kappa
ResNet34 (ImageNet)	96.5	82.9	82.4	88.8
ResNet34 (Fundus)	97.0	85.0	84.7	90.2
DLI [13]	–	82.5	80.3	89.5
CANet [10]	–	83.2	81.3	90.0
GREEN-ResNet50 [11]	–	84.4	83.6	90.8
GREEN-SE-ResNext50 [11]	–	**85.7**	85.2	91.2
MIL-VT (Proposed)	**97.9**	85.5	**85.3**	**92.0**

Table.1 Comparison with state-of-the-art methods on APTOS2019

Method	AUC	Acc	F1	Recall	Precision
ResNet34 (ImageNet)	93.5	87.8	92.6	92.4	92.1
ResNet34 (Fundus)	94.7	89.1	93.0	91.4	94.5
MIL-VT (Proposed)	**95.9**	**91.1**	**94.4**	**93.7**	**95.0**

Table.2 Comparison with state-of-the-art methods on RFMiD2020

3.2 Ablation study

Model	Combination		DR Grading			
	Pre-train	MIL	AUC	Acc	F1	Kappa
VT (ImageNet)			96.7	82.3	81.5	89.0
VT (Fundus)	✓		97.5	84.6	83.8	91.1
MIL-VT (Fundus)	✓	✓	**97.9**	**85.5**	**85.3**	**92.0**

Table 3. The ablation study results of MIL-VT on APTOS2019

Model	Combination		Disease Classification				
	Pre-train	MIL	AUC	Acc	F1	Recall	Precision
VT (ImageNet)			93.0	86.5	91.2	89.1	93.4
VT (Fundus)	✓		94.5	88.5	92.6	90.4	94.8
MIL-VT (Fundus)	✓	✓	**95.9**	**91.1**	**94.4**	**93.7**	**95.0**

Table 4. The ablation study results of MIL-VT on RFMiD2020

（MICCAI 2021）CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation

—— Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia

1. Motivation

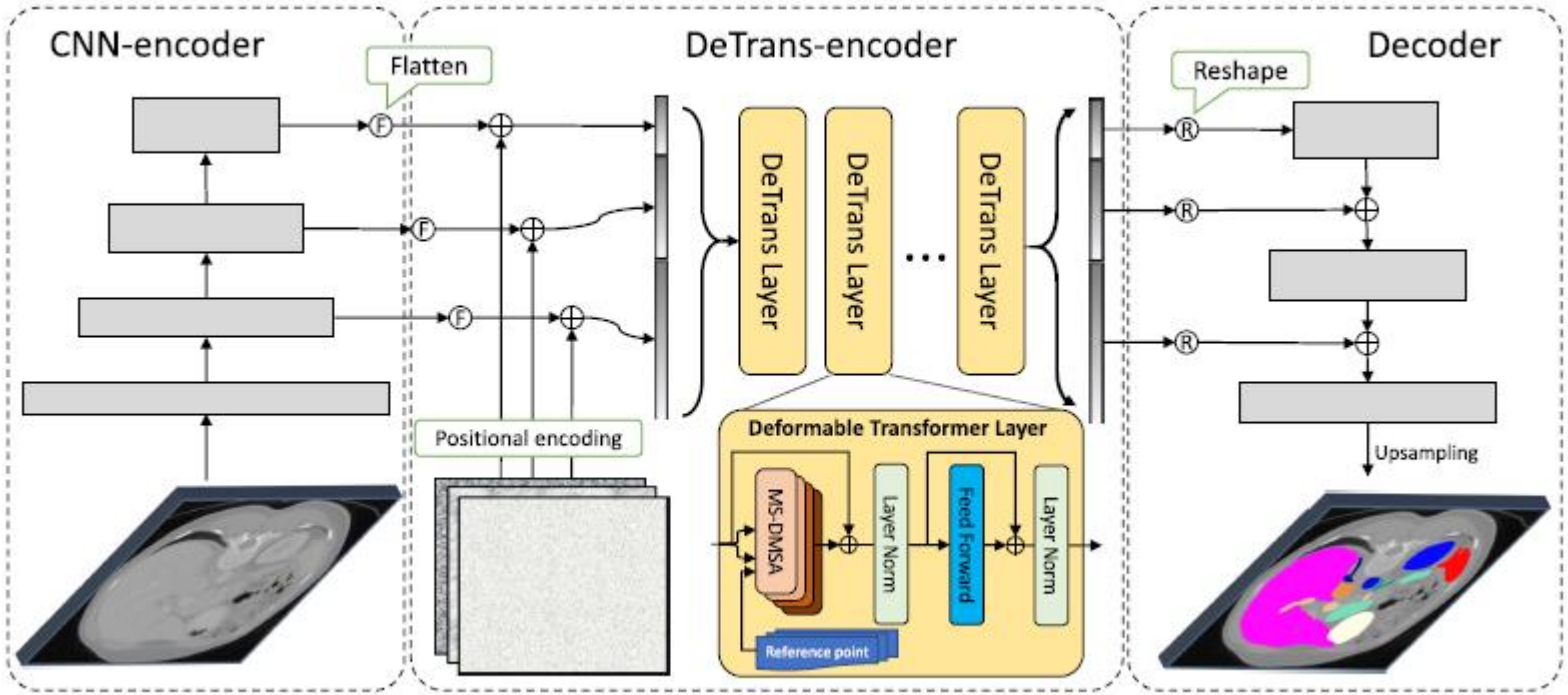
- Apply transformer into 3D medical image segmentation.
- Combine CNN with Transformer.
- Reduce the spatial and computational complexity.

2. Dataset

2.1 Multi-Atlas Labeling **B**eyond the **C**ranial **V**ault (BCV) dataset

Contains 30 labeled CT scans for automated segmentation of 11 abdominal organs (腹部器官), including the spleen (Sp), kidney (Ki), gallbladder (Gb),esophagus (Es), liver (Li), stomach (St), aorta (Ao), inferior vena cava (IVC), portal vein and splenic vein (PSV), pancreas (Pa), and adrenal gland (AG).

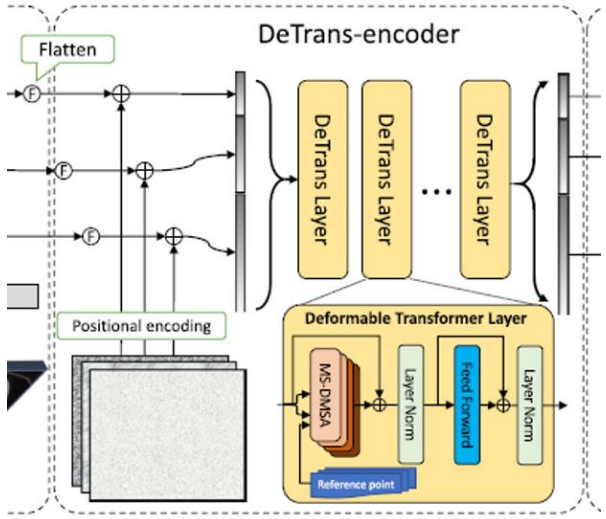
3. Framework


CNN-encoder
Flatten
DeTrans-encoder
DeTrans Layer
DeTrans Layer
DeTrans Layer
Positional encoding
Deformable Transformer Layer
MS-DMSA
Layer Norm
Feed Forward
Layer Norm
Reference point
Decoder
Reshape
Upsampling

3.1 CNN encoder

Three stage with three, three and two 3D residual blocks respectively.
The input image *x* is a combination of slice, with dimension (H*W*D)

3.2 DeTrans-Encoder



3.2.1 Input-to-Sequence transformation

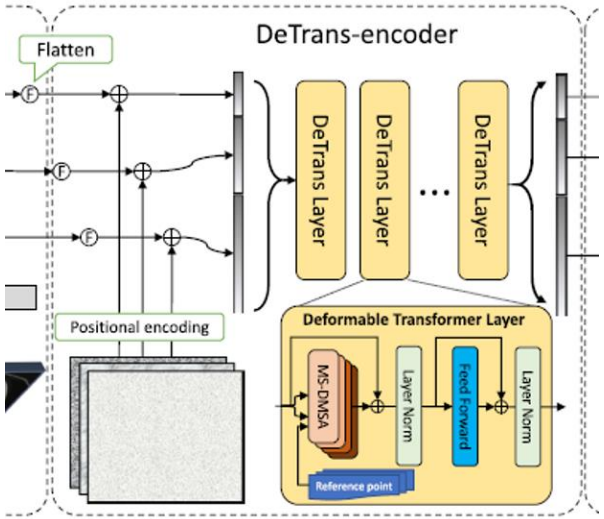
1. Feature maps extracted by CNN will be flattened to 1-D sequence f_l
2. To make up for the spatial information loss, using 3D positional encoding.
   '#' represents {D,H,W};

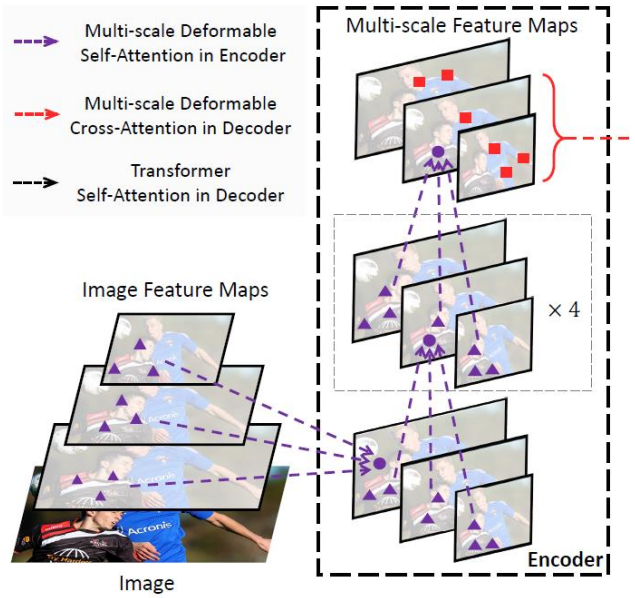
$$\begin{cases} PE_{\#}(pos, 2k) = sin(pos \cdot \upsilon) \\ PE_{\#}(pos, 2k+1) = cos(pos \cdot \upsilon) \end{cases} \qquad \upsilon = 1/10000^{2k/\frac{C}{3}}$$

PE_D, PE_H, PE_W are concatenated as P_l. Then combine P_l and f_l via element-wise summation

3.2 DeTrans-Encoder

3.2.2 MS-DMSA Layer (Multi-scale Deformable MSA)





Normal MSA:

$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \big],$$

Deformable MSA:

$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big],$$

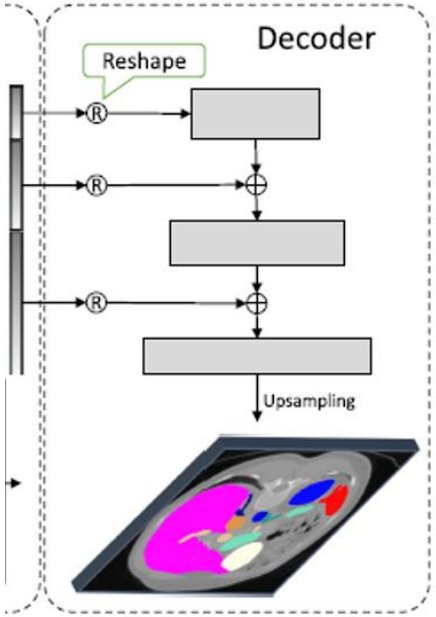
$\boldsymbol{A_{mqk}}$: Attention matrix for attention head m
$\boldsymbol{x_k}$: H*W Value feature
$\boldsymbol{x(p_q + \Delta p_{mqk})}$: K sample points, $p_q + \Delta p_{mqk}$ represents the position. $p_q$is the reference point which z_q corresponding to, and Δp_{mqk} is a offset via linear projection of z_q.

MS-DMSA:

$$\text{head}_i = \sum_{l}^{L} \sum_{k}^{K} \Lambda(\boldsymbol{z}_q)_{ilqk} \cdot \Psi(\boldsymbol{f}_l)(\sigma_l(\hat{\boldsymbol{p}}_q) + \Delta_{\boldsymbol{p}_{ilqk}})$$

$$\text{MS} - \text{DMSA}(\boldsymbol{z}_q, \{\boldsymbol{f}_l\}_{l=1}^{L}) = \Phi(\text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_H))$$

3.2.3 Decoder



1. The output sequence of each scale is reshaped into feature maps.
2. Similar to U-net, upsampling feature maps and concat them.
3. Refine them using a 3D residual block.

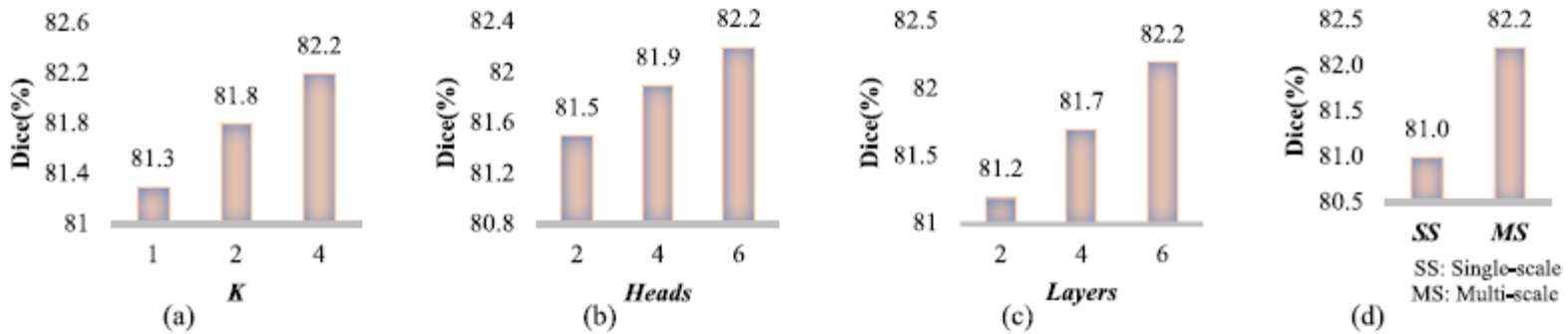
4. Experiments

Table 1. Dice and HD scores of our CoTr and several competing methods on the BCV test set. **CoTr*** and **CoTr**† are two variants of CoTr with small CNN-encoders.

Methods	Param (M)	Dice scores of different organs											Ave Dice	Ave HD
		Sp	Ki	Gb	Es	Li	St	Ao	IVC	PSV	Pa	AG		
SETR (ViT-B/16-rand) [27]	100.5	95.2	92.3	55.6	71.3	96.2	80.2	89.7	83.9	68.9	68.7	60.5	78.4	7.47
SETR (ViT-B/16-pre) [27]	100.5	94.8	91.7	55.2	70.9	96.2	76.9	89.3	82.4	69.6	70.7	58.7	77.8	8.47
CoTr w/o CNN-encoder	21.9	95.2	92.8	59.2	72.2	96.3	81.2	89.9	85.1	71.9	73.3	61.0	79.8	7.23
CoTr w/o DeTrans	32.6	96.0	92.6	63.8	77.9	97.0	83.6	90.8	87.8	76.7	81.2	72.6	83.6	5.18
APSS [5]	45.5	96.5	93.8	65.6	78.1	97.1	84.0	91.1	87.9	77.0	82.6	73.9	84.3	4.85
PP [26]	33.9	96.1	93.1	64.3	77.4	97.0	85.3	90.8	87.4	77.2	81.9	72.8	83.9	5.10
Non-local [20]	32.8	96.3	93.7	64.6	77.9	97.1	84.1	90.8	87.7	77.2	82.1	73.3	84.1	4.70
TransUnet [4]	43.5	95.9	93.7	63.1	77.8	97.0	86.2	91.0	87.8	77.8	81.6	73.9	84.2	4.77
CoTr*	27.9	96.4	94.0	66.2	76.4	97.0	84.2	90.3	87.6	76.3	80.8	72.9	83.8	5.04
CoTr†	36.9	96.2	93.8	66.5	78.6	97.1	86.9	90.8	87.8	77.7	82.8	73.2	84.7	4.39
CoTr	41.9	96.3	93.9	66.6	78.0	97.1	88.2	91.2	88.0	78.1	83.1	74.1	**85.0**	**4.01**

CoTr with different CNN-encoders: one 3D residual block ; Two 3D residual block; 3, 3, 2;
Pure CNNs: CoTr w/o DeTrans ; Atrous Spatial Pyramid Pooling (ASPP) ; pyramid parsing (PP) ; Nonlocal module
Hybrid CNN-Transformer Encoder: 3D TransUNet

4.2 Parameter Settings



Number of key points, Heads, DeTrans layers; Multi-scale