Medical Vision Seminar

Yujin Tang

21.09.22

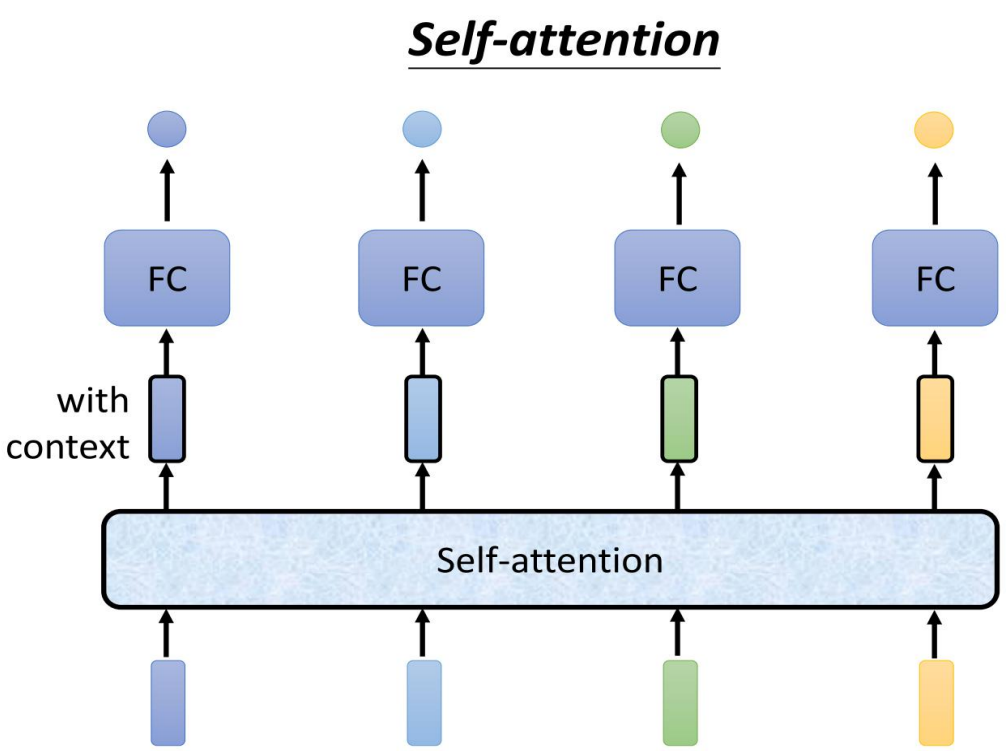
Transformer for Medical Image Segmetation

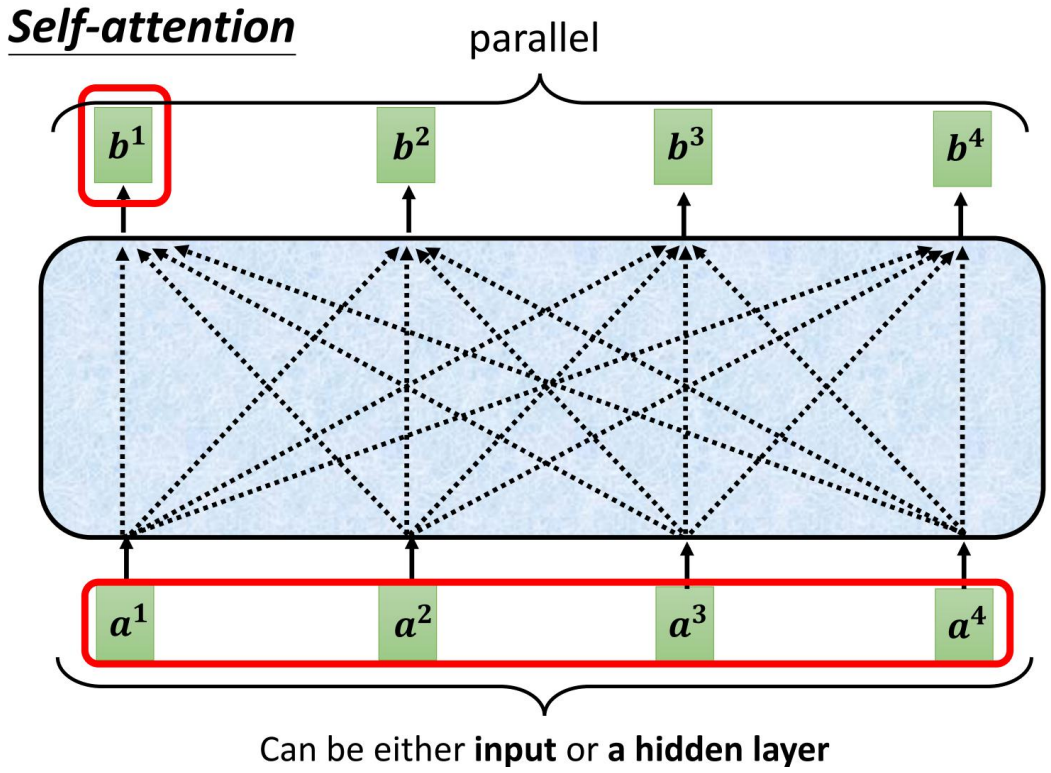
Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Zhou, H. Y., Guo, J., Zhang, Y., Yu, L., Wang, L., & Yu, Y. (2021). nnFormer: Interleaved Transformer for Volumetric Segmentation. arXiv preprint arXiv:2109.03201.

Background--Self Attention

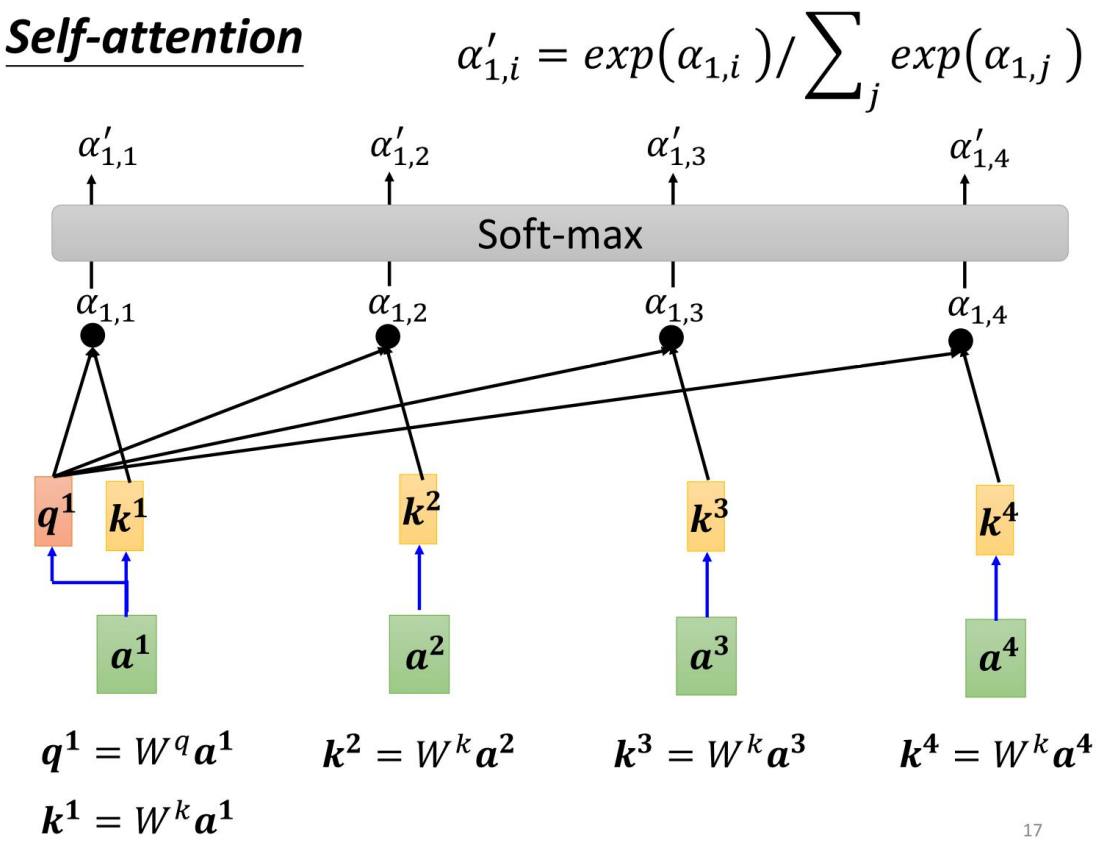
- seq2seq(Model decides the number of labels itself)

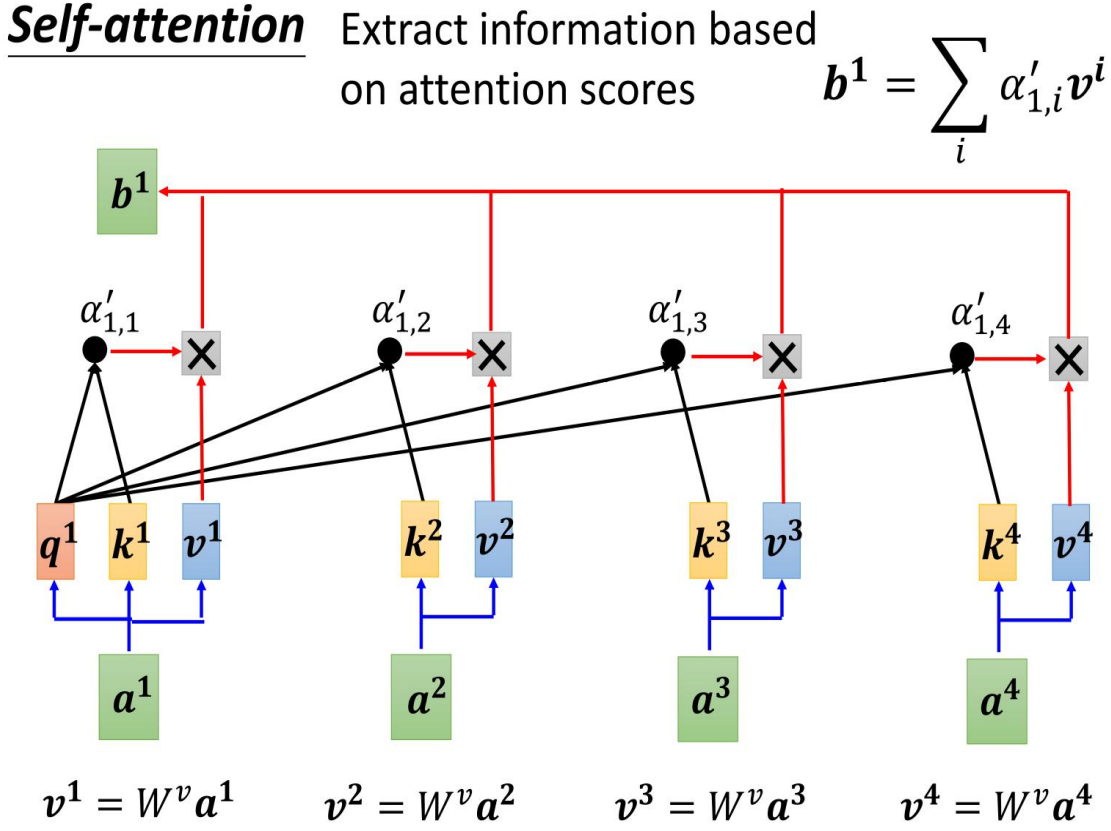




Background--Self Attention

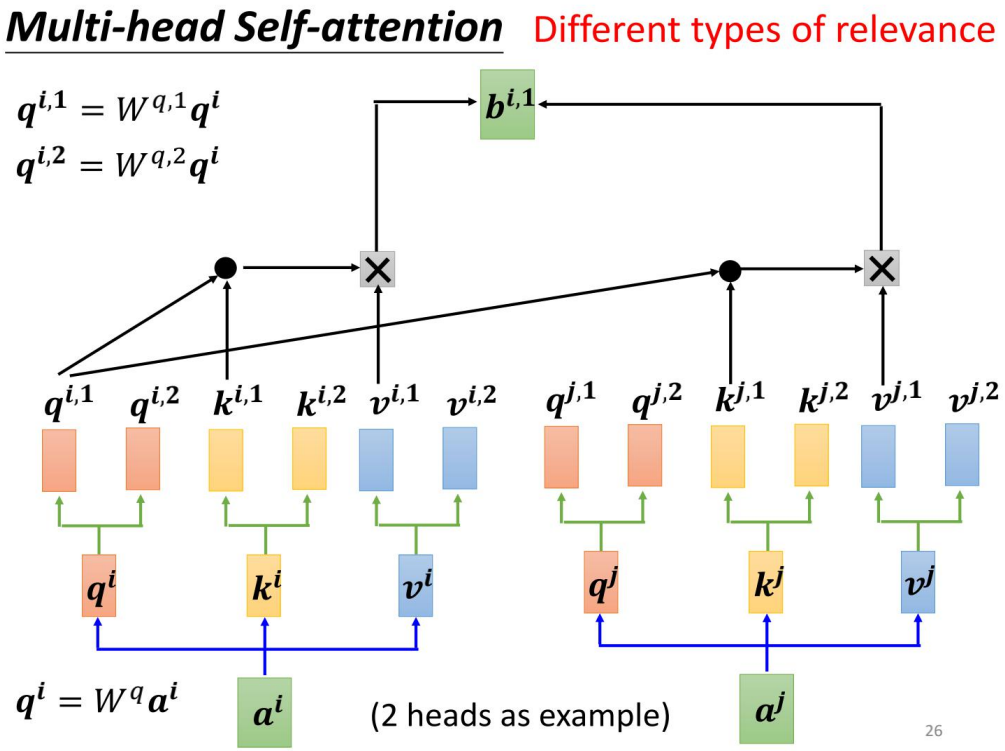
- (Q,K,V)--learn w^q w^k w^v





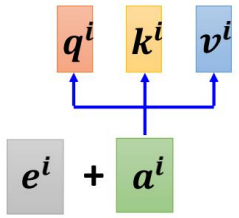
Background--Self Attention

- MSA(Multi-head Self Attention)
- Positional Enconding

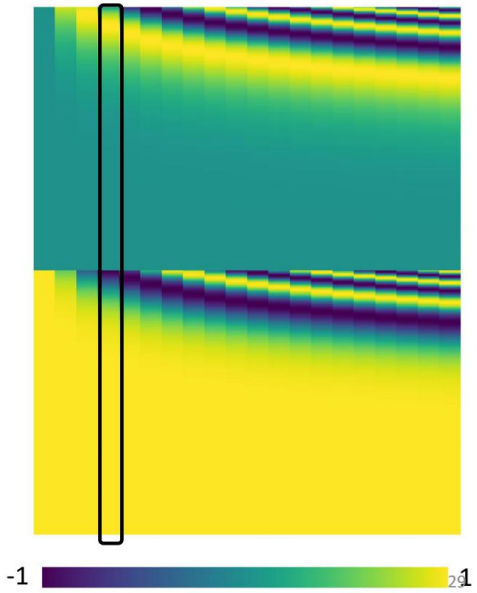


Positional Encoding

- No position information in self-attention.
- Each position has a unique positional vector e^i
- **hand-crafted**
- **learned from data**

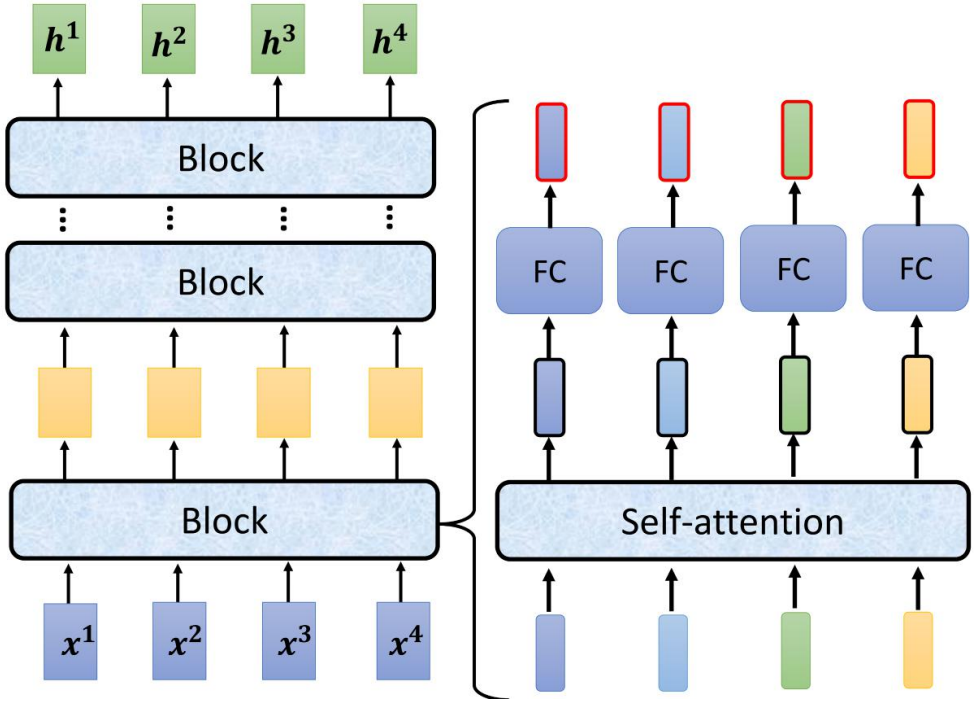


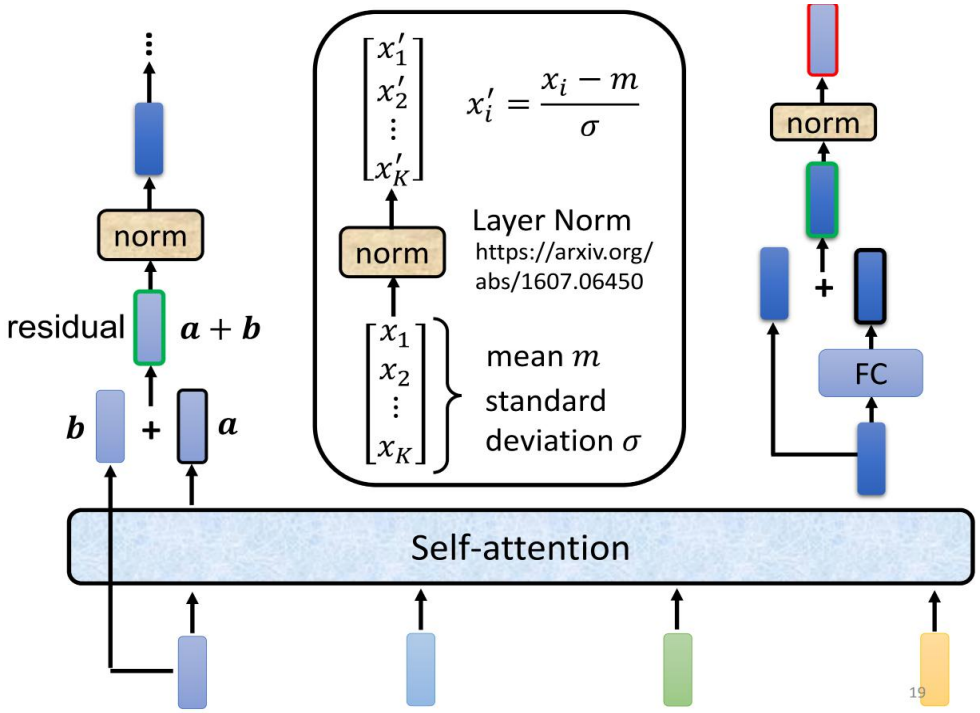
Each column represents a positional vector e^i



Background--Transformer

- Residual+Layer Normalization





TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

Jieneng Chen[1], Yongyi Lu[1], Qihang Yu[1], Xiangde Luo[2],
Ehsan Adeli[3], Yan Wang[4], Le Lu[5], Alan L. Yuille[1], and Yuyin Zhou[3]

[1]Johns Hopkins University
[2]University of Electronic Science and Technology of China
[3]Stanford University
[4] East China Normal University
[5]PAII Inc.

Introduction

- Motivation:
  - CNN-based approaches generally exhibit limitations for modeling explicit long-range relation.
  - Transformers are powerful at modeling global contexts and demonstrate superior transferability for downstream tasks.

- Contribution: The first study which explores the potential of transformers in the context of medical image segmentation.

Method

Patchsize:P*P
Sequence Length:
H*W/P*P

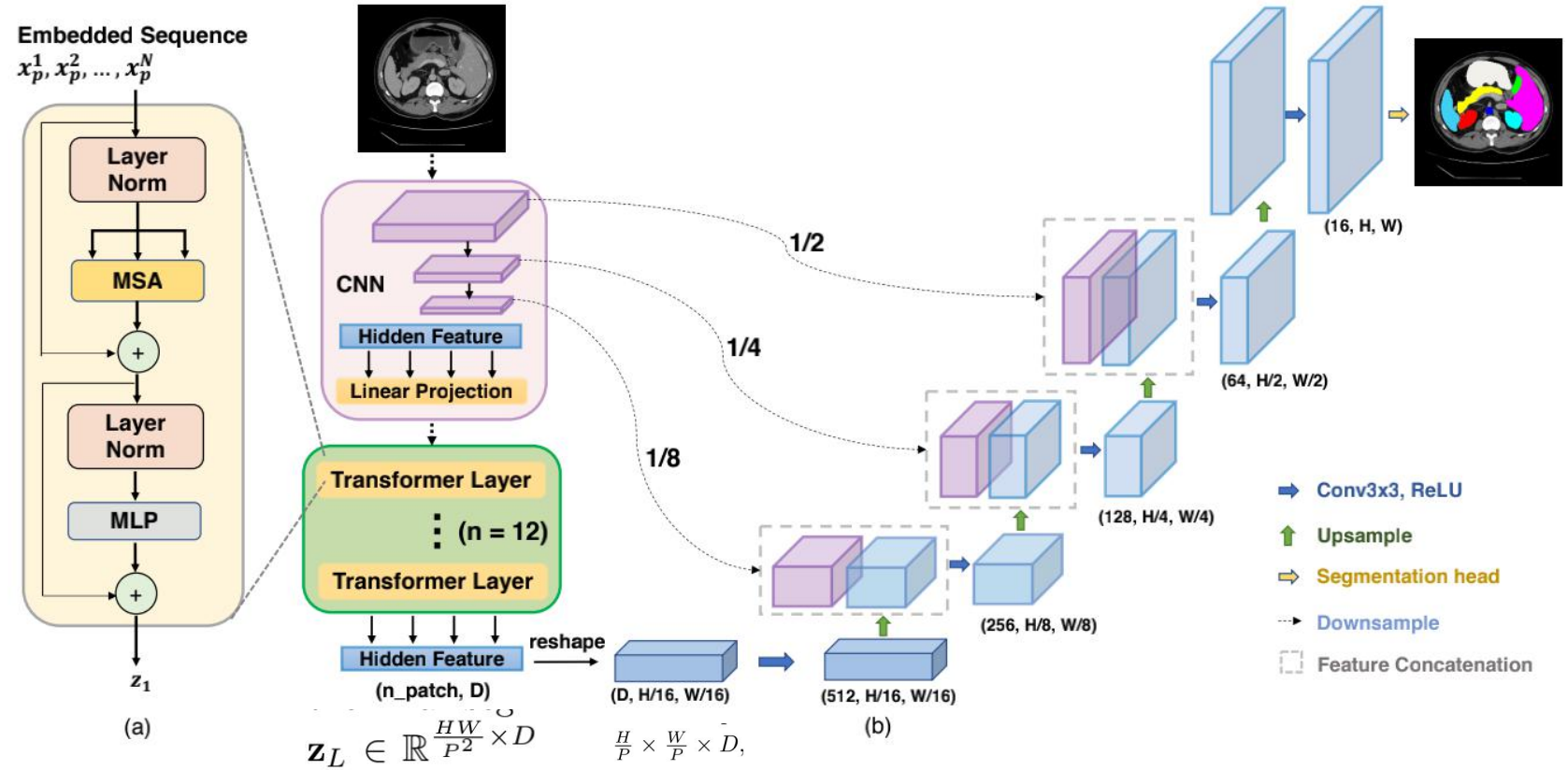


Fig. 1: Overview of the framework. (a) schematic of the Transformer layer; (b) architecture of the proposed TransUNet.

Method

$$\mathbf{z}_0 = [\mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\cdots;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \tag{1}$$

$$\mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D} \qquad \mathbf{E}_{pos} \in \mathbb{R}^{N\times D}$$

$$\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \tag{2}$$

$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \tag{3}$$

Result

Table 1: Comparison on the Synapse multi-organ CT dataset (average dice score % and average hausdorff distance in mm, and dice score % for each organ).

Framework		**Average**		Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Encoder	Decoder	DSC ↑	HD ↓								
V-Net [9]		68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [5]		69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50	U-Net [12]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50	AttnUNet [13]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
ViT [4]	None	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
ViT [4]	CUP	67.86	36.11	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-ViT [4]	CUP	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet		**77.48**	**31.69**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62

Result--Ablation Study

- number of skip connections
- input resolution
- patch size and sequence length
- model scale

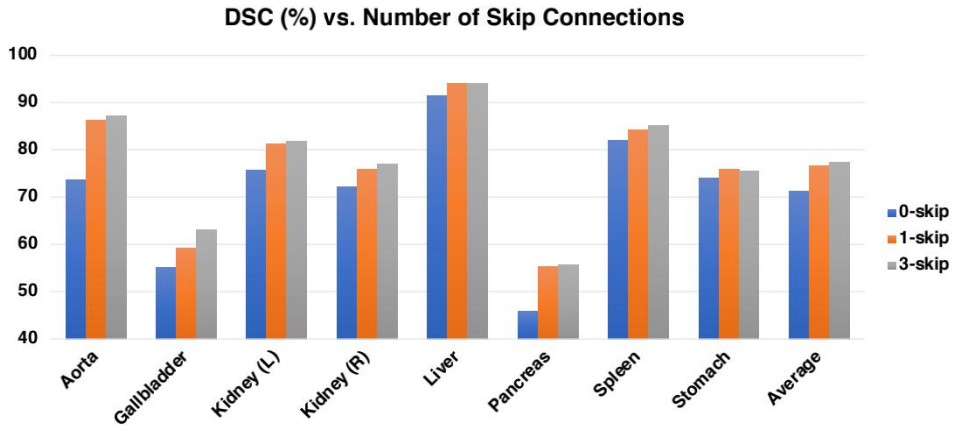


Fig. 2: Ablation study on the number of skip-connections in TransUNet.

Table 2: Ablation study on the influence of input resolution.

Resolution	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
224	**77.48**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
512	**84.36**	90.68	71.99	86.04	83.71	95.54	73.96	88.80	84.20

Table 3: Ablation study on the patch size and the sequence length.

Patch size	Seq_length	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
32	49	76.99	86.66	63.06	81.61	79.18	94.21	51.66	85.38	74.17
16	196	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
8	784	**77.83**	86.92	58.31	81.51	76.40	93.81	58.09	87.92	79.68

Table 4: Ablation study on the model scale.

Model scale	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Base	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Large	78.52	87.42	63.92	82.17	80.19	94.47	57.64	87.42	74.90

hidden size D, number of layers, MLP size, number of heads:
12, 768, 3072, 12
24, 1024, 4096, 16

Result--Visualization

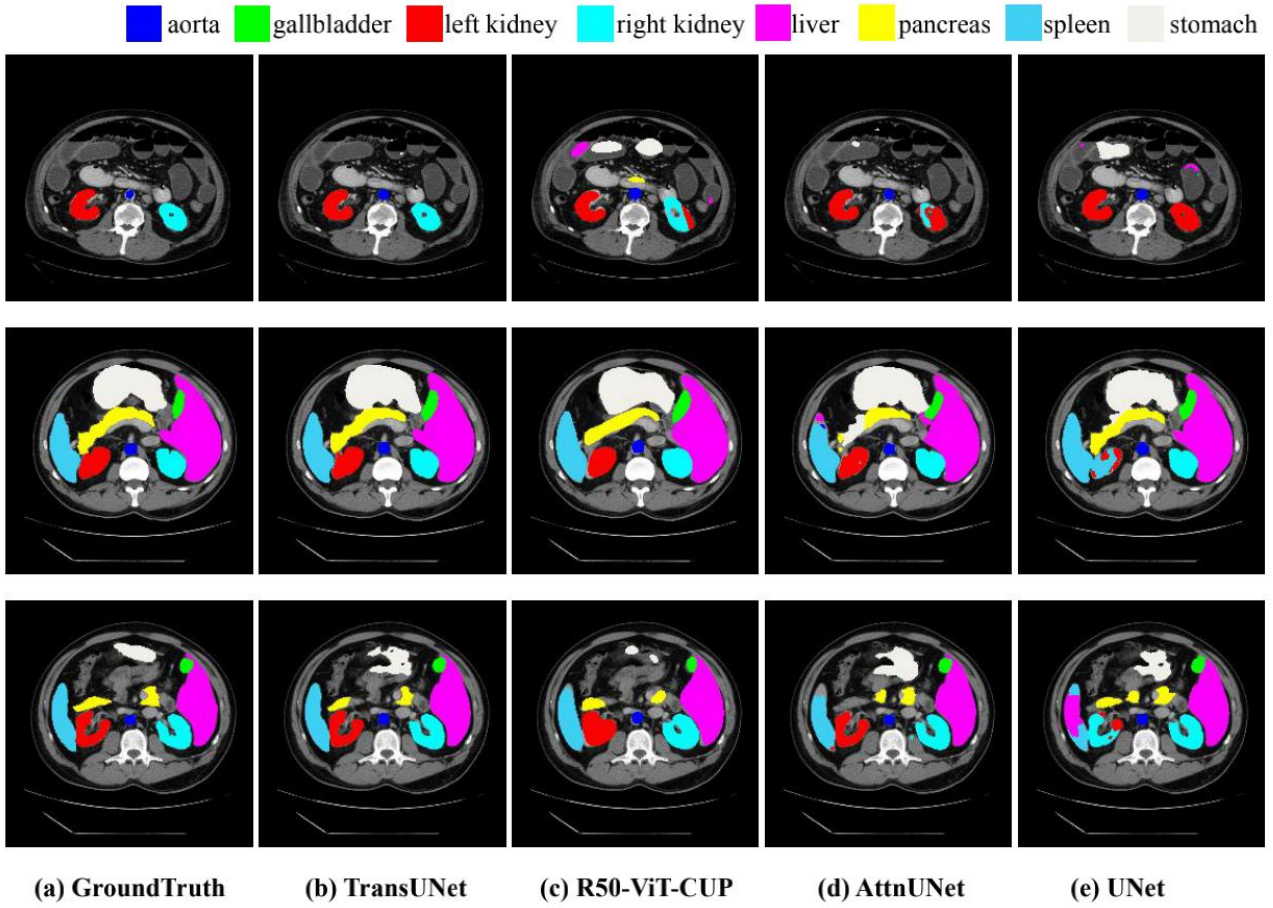


Fig. 3: Qualitative comparison of different approaches by visualization. From left to right: (a) Ground Truth, (b) TransUNet, (c) R50-ViT-CUP, (d) R50-AttnUNet, (e) R50-U-Net. Our method predicts less false positive and keep finer information.

nnFormer: Interleaved Transformer for Volumetric Segmentation

Hong-Yu Zhou[⋆1,2], Jiansen Guo[⋆1], Yinghao Zhang[⋆1], Lequan Yu[3], Liansheng Wang[1], and Yizhou Yu[2]

[1]Department of Computer Science, Xiamen University

[2]Department of Computer Science, The University of Hong Kong

[3]Department of Statistics and Actuarial Science, The University of Hong Kong

whuzhouhongyu@gmail.com, {jsguo,zhangyinghao}@stu.xmu.edu.cn, lqyu@hku.hk, lswang@xmu.edu.cn, yizhouy@acm.org

Introduction

- Motivation:
  - TransUNet treats convnets as feature extractors and uses transformers to help encode the global context.
  - SwinUNet: Pure transformer
  - Both convnets and transformers did not explore how to appropriately combine convolution and self-attention for building an optimal medical segmentation network.
- Contribution:
  - Hybrid stem where convolution and self-attention are interleaved to give full play to their strengths.
  - Propose a computational-efficient way to capture inter-slice dependencies. 98% on Synapse and 99.5% on ACDC.

Method

$$\mathcal{X}_t \in \mathcal{R}^{L \times C}$$

$$\hat{\mathcal{X}}_t \in \mathbf{R}^{N_V \times N_T \times C} \qquad \{S_H,\ S_W,\ S_D\}$$

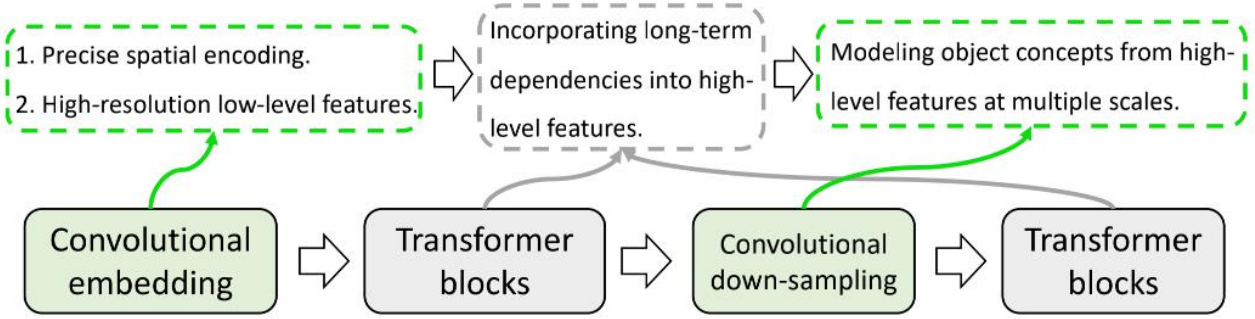
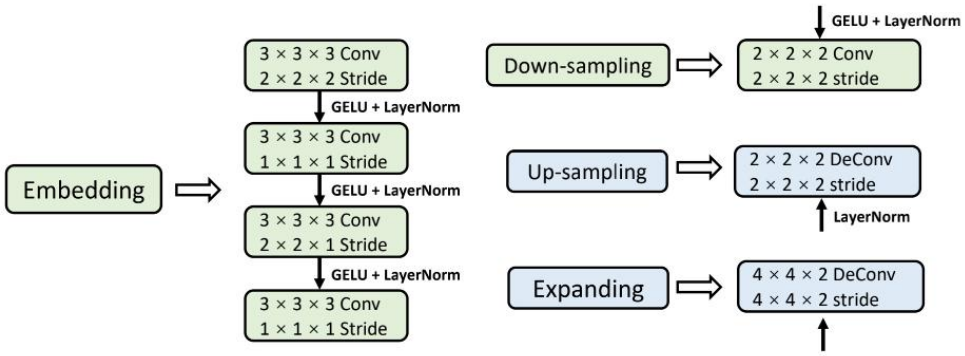
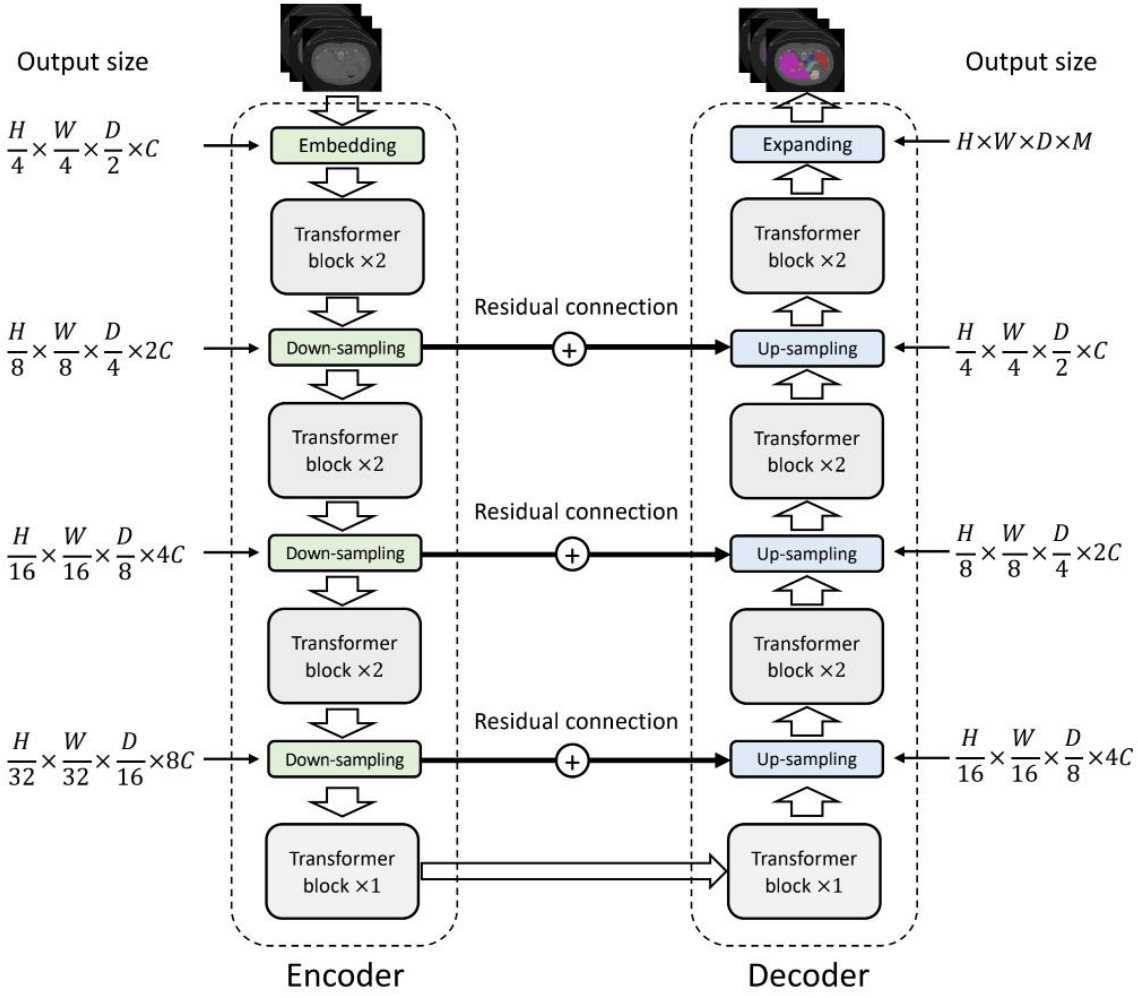


Fig. 1: Overview of the interleaved stem used in the encoder of nnFormer.

b.



a.



Method

$$\hat{\mathcal{X}}_t^l = \text{V-MSA}\left(\text{LayerNorm}\left(\mathcal{X}_t^{l-1}\right)\right) + \mathcal{X}_t^{l-1},$$
$$\mathcal{X}_t^l = \text{MLP}\left(\text{LayerNorm}\left(\hat{\mathcal{X}}_t^l\right)\right) + \hat{\mathcal{X}}_t^l,$$
$$\hat{\mathcal{X}}_t^{l+1} = \text{SV-MSA}\left(\text{LayerNorm}\left(\mathcal{X}_t^l\right)\right) + \mathcal{X}_t^l,$$
$$\mathcal{X}_t^{l+1} = \text{MLP}\left(\text{LayerNorm}\left(\hat{\mathcal{X}}_t^{l+1}\right)\right) + \hat{\mathcal{X}}_t^{l+1}.$$

$$\{S_H,\ S_W,\ S_D\}$$

$$\left\lfloor \frac{S_H}{2} \right\rfloor,\ \left\lfloor \frac{S_W}{2} \right\rfloor,\ \left\lfloor \frac{S_D}{2} \right\rfloor$$

$$\Omega(\text{V-MSA}) = 4hwdC^2 + 2S_H S_W S_D hwdC.$$

$$\Omega(\text{MSA}) = 4hwdC^2 + 2(hwd)^2 C.$$

$$\text{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V,$$

Synapse:[4,4,4]
ACDC:[5,5,3]

Dataset

Synapse for multi-organ CT segmentation. This dataset includes 30 cases of abdominal CT scans. Following the split used in [6], 18 cases are extracted to build the training set while the rest 12 cases are used for testing. We report the model performance evaluated with the average Dice Similarity Coefficient (DSC) on 8 abdominal organs, which are aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach.

ACDC for automated cardiac diagnosis. ACDC involves 100 patients, with the cavity of the right ventricle, the myocardium of the left ventricle and the cavity of the left ventricle to be segmented. Each case's labels involve left ventricle (LV), right ventricle (RV) and myocardium (MYO). The dataset is split into 70 training samples, 10 validation samples and 20 testing samples.

Result

Table 3: Experiments on ACDC (dice score in %). Best results are bolded.

Methods	Average	RV	Myo	LV
R50-U-Net [28]	87.55	87.10	80.63	94.92
R50-Attn UNet [29]	86.75	87.58	79.20	93.47
VIT-CUP [8]	81.45	81.46	70.71	92.18
R50-VIT-CUP [8]	87.57	86.07	81.88	94.75
CBAM [36]	87.30	87.70	82.10	92.20
ResUNet [10]	86.90	86.20	82.50	92.20
Dual-Attn [10]	87.00	86.40	82.30	92.40
UTNET [10]	88.30	88.20	83.50	93.10
TransUNet [6]	89.71	88.86	84.54	95.73
SwinUNet [3]	90.00	88.55	85.62	**95.83**
LeViT-UNet-384s [40]	90.32	89.55	87.64	93.76
nnUNet (3D) [12]	91.59	**90.25**	89.10	95.41
nnFormer	**91.78**	90.22	**89.53**	95.59

Result--Visualization

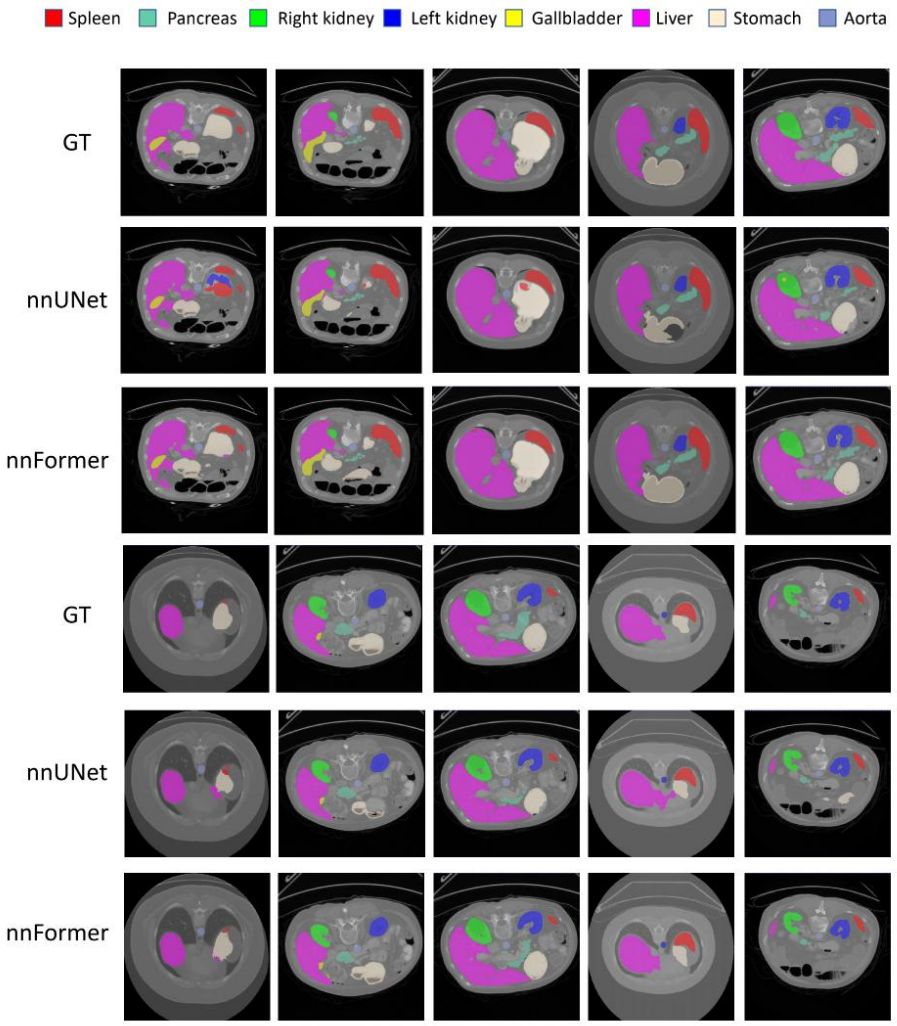


Fig. 3: Segmentation results of some hard samples on Synapse.

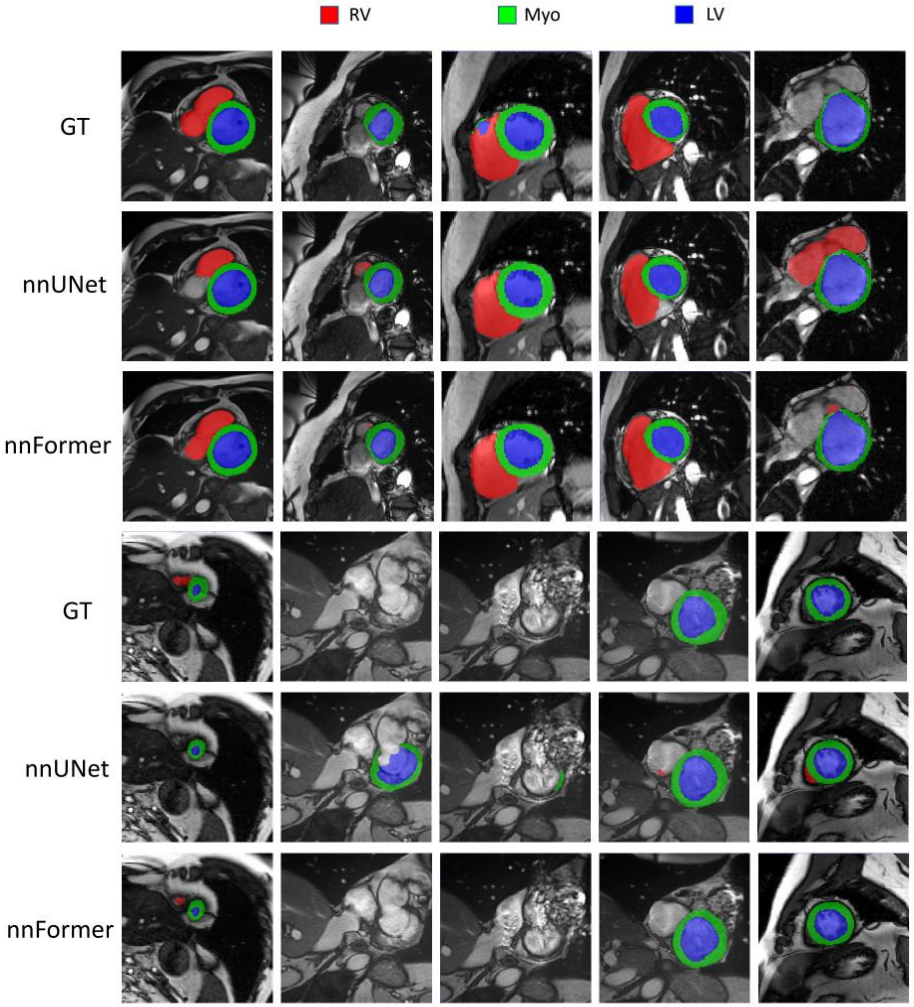


Fig. 4: Segmentation results of some hard samples on ACDC.

Result--Ablation Study

Table 4: Investigation of the embedding block on Synapse. **Patch-wise convolution** consists of only one convolutional layer with large kernel size and stride.

	Average	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
Patch-wise convolution	84.63	88.84	65.33	86.97	85.98	95.58	77.30	91.83	85.15
Ours	**87.40**	**92.04**	**71.09**	**87.64**	**87.34**	**96.53**	**82.49**	**92.91**	**89.17**

Table 5: Investigation of the convolutional down-sampling blocks.

	Average	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
Neighboring concatenation	84.30	88.00	67.60	87.52	87.38	95.31	80.63	85.29	82.69
Ours	**87.40**	**92.04**	**71.09**	**87.64**	**87.34**	**96.53**	**82.49**	**92.91**	**89.17**

Table 6: Investigation of adding more transformer blocks.

	Average	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
More transformer blocks	85.98	89.02	**71.74**	86.76	87.06	96.37	82.30	89.04	85.51
Ours	**87.40**	**92.04**	71.09	**87.64**	**87.34**	**96.53**	**82.49**	**92.91**	**89.17**

Table 7: Benefits of using pre-trained weights on natural images.

	Average	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
No pre-training	84.34	90.15	69.00	86.34	87.48	95.93	80.97	85.23	79.67
Ours	**87.40**	**92.04**	71.09	**87.64**	**87.34**	**96.53**	**82.49**	**92.91**	**89.17**

- large convolution kernel VS successive small kernels. one [4,4,2]
- neighboring contenation VS down-sampling blocks
- 2 transformer blocks VS 1 transformer block
- no pre-training VS pretraining

Thank You!