Medical Vision Seminar

Yujin Tang

2021.12.08

- (MICCAI2021) Spine-Transformers: Vertebra Detection and Localization in Arbitrary Field-of-View Spine CT with Transformers

- (MICCAI2021) Multi-compound Transformer for Accurate Biomedical Image Segmentation

Spine-Transformers: Vertebra Detection and Localization in Arbitrary Field-of-View Spine CT with Transformers

Rong Tao and Guoyan Zheng(✉)

Institute of Medical Robotics, School of Biomedical Engineering,
Shanghai Jiao Tong University, No. 800, Dongchuan Road,
Shanghai 200240, China
guoyan.zheng@sjtu.edu.cn

Introduction

- Input：3D arbitrary FOV（任意视野）Spine CT
- Task：vertebra detection and localization
- Previous Study:

✓required complicated steps to handle arbitrary FOV issue, e.g.machine learning + hidden Markov model, 3D FCN on down-scaled image + hidden Markov model as post processing

✓were difficult to be applied to scans with arbitrary FOV due to the use of pre-defined adjacency matrix, e.g. Graph Convolutional Network-based method

Motivation

- DETR: first end-to-end object detector
- Remove need for NMS and anchor generation
- Transformer-based methods have never been applied to object detection for 3D images
- Hungarian algorithm, designed a loss function to find a bipartite marching between GT and prediction, and requires permutation-invariance.

Contribution

- The first to apply a transformers-based model to 3D medical image detection task
- Formulate the automatic vertebra detection problem as a direct one-to-one set prediction problem and introduce a new one-to-one set-based global loss, can directly output all vertebrae in parallel
- Introduce a novel inscribed sphere-based object detector which can handle volume orientation variation better than the regular box-based detector.

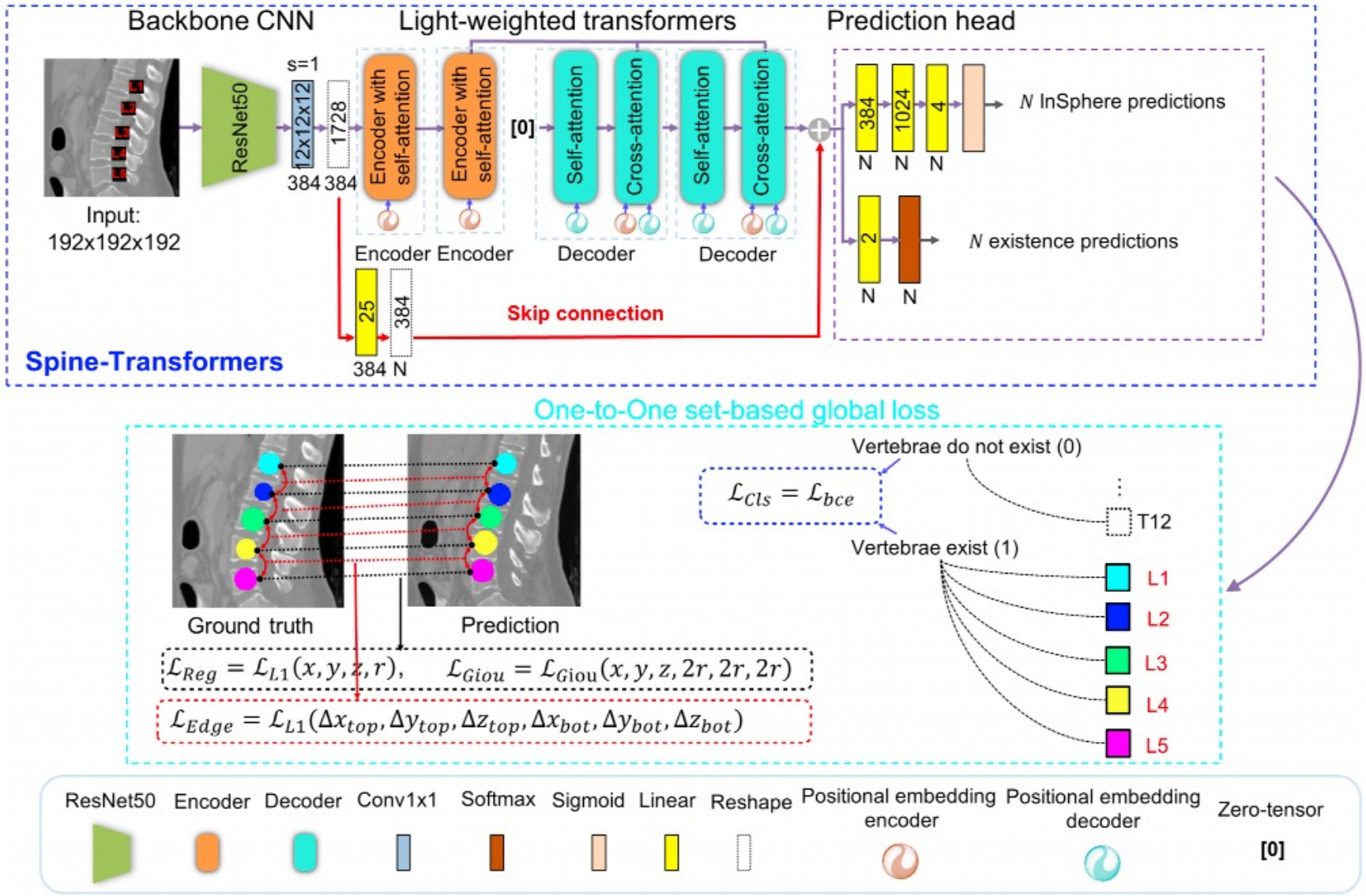
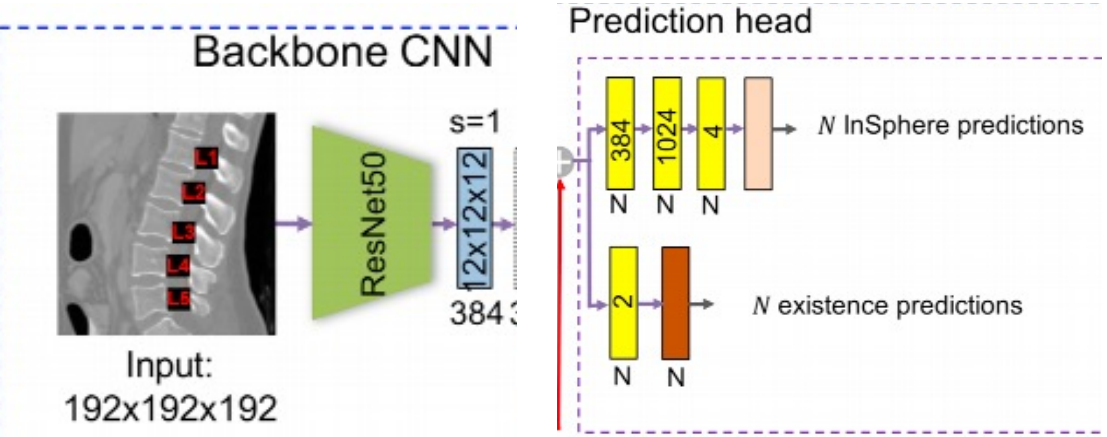
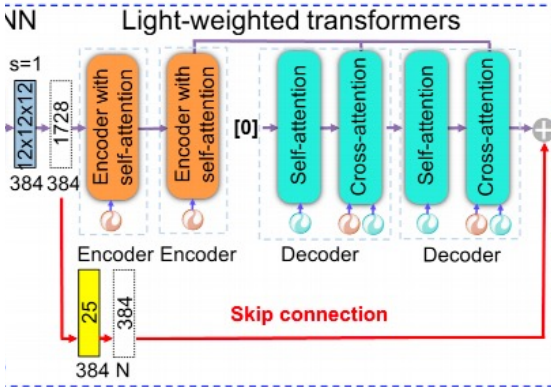


Fig. 1. A schematic illustration of the Spine-Transformers and on how the losses are computed. See text for detailed explanation.

Methods-Block



✓ Input: fixed-size patches
✓ N: maximum number of vertebra levels found in GT
✓ Backbone: ResNet50
Output: H/16*W/16*D/16*2048

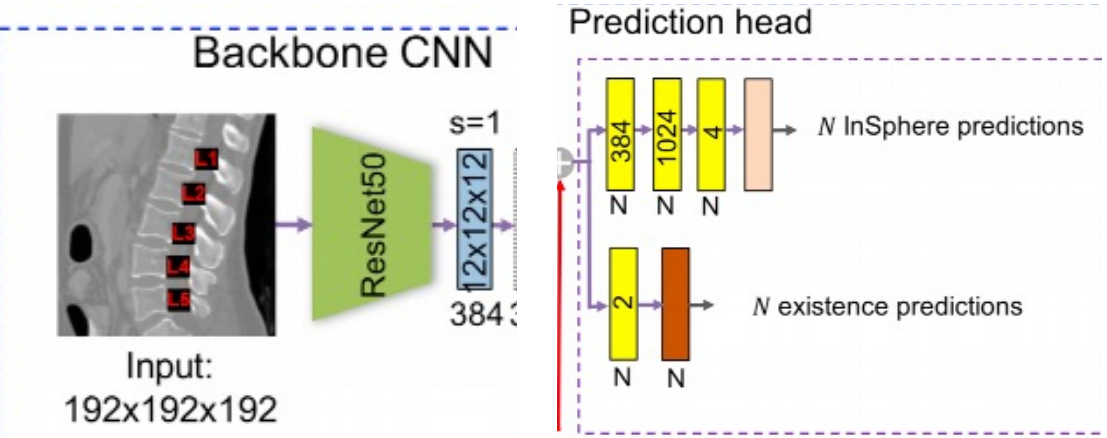


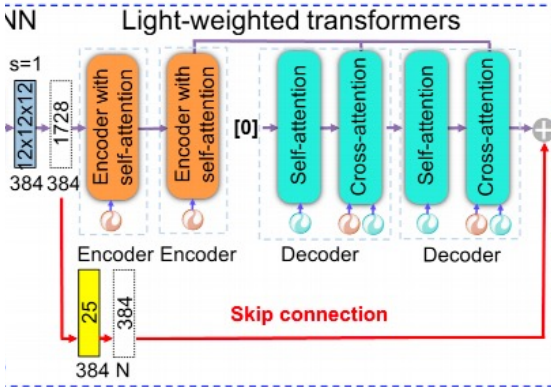
✓ Transformer Encoder:
1*1*1 convolution: reduce channel dimension from C to 3x d_{hidden} , 128, learnable encoder positional embedding

✓ Transformer Decoder:
Yield N predictions, add learnable decoder positional embedding to each layer of decoder

✓ Skip connection:
Alleviate transformer-based model's data hungry and extra-long time convergency problem

Methods-Block





✓ Light-weighted Architecture Design:

Increasing resolution of bottom feature map-> GPU Memory

Reducing resolution of bottom feature map-> can't faciliate non-local global context information

Two –layer encoder and two-layer decoder

✓ Localization Refinement:

Use 3D CNN-based landmark regression to improve localization accuracy. Crop a sub-volume of fixed size around detected center, as input of 3D CNN for refinement

Methods

- Classic box detectors are not sufficient, as they are not rotational invariant.
- InSphere detector

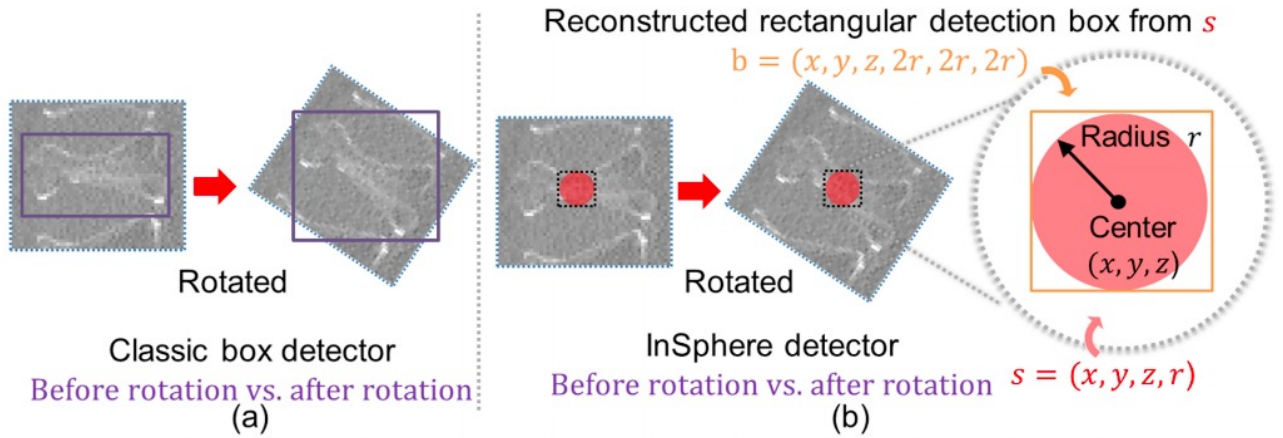
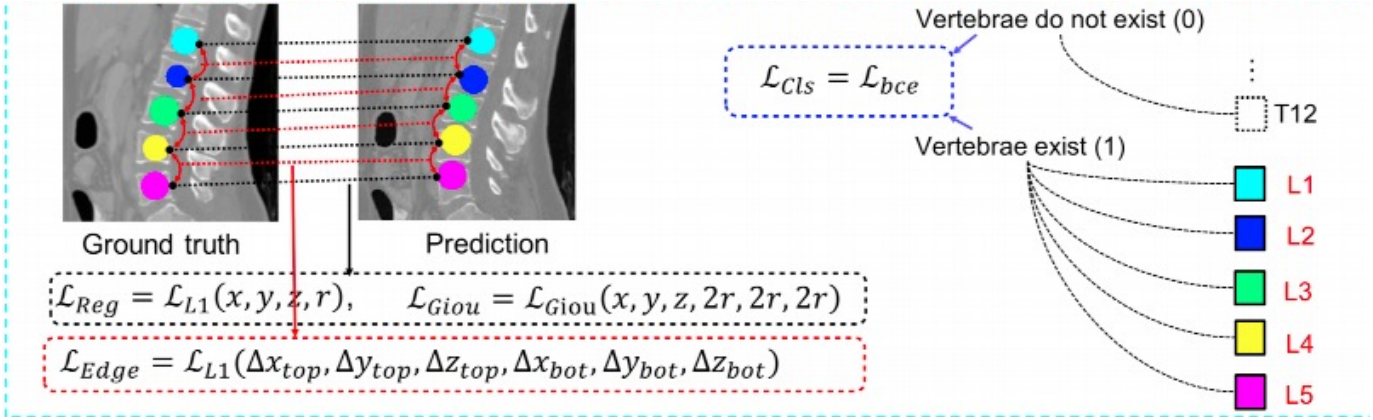


Fig. 2. A comparison of box detector (a) with InSphere detector (b). InSphere detection is not sensitive to vertebral orientation.

Methods-Loss



- ✓ GT label: v_i=(c_i, s_i) c_i: ϵ{0,1} s_i: [x_i, y_i, z, r_i]
- ✓ Binary cross-entropy loss for c_i

$$L_{Cls} = -\frac{1}{N} \cdot \sum_{i=1}^{N} (c_i log(\hat{c}_i) + (1 - c_i) log(1 - \hat{c}_i)) \quad (1)$$

- ✓ Regression loss for all n vertebrae s_i

$$L_{Reg} = \frac{1}{n} \cdot \sum_{i=1}^{n} |(x_i - \hat{x}_i)| + |(y_i - \hat{y}_i)| + |(z_i - \hat{z}_i)| + |(r_i - \hat{r}_i)| \quad (2)$$

- ✓ Edge loss for top and bottom neighborhood vertebral center

$$L_{Edges} = \frac{1}{n} \cdot \sum_{i=1}^{n} (|(edge_{i,top} - \hat{edge}_{i,top})| + |(edge_{i,bottom} - \hat{edge}_{i,bottom}|) \quad (3)$$

- ✓ Area loss GIoU, cube b_i=(x_i, y_i, z_i,2 r_i,2 r_i,2 r_i) from s_i

$$L_{Giou} = \frac{1}{n} \cdot \sum_{i=1}^{n} \{1 - (\frac{b_i \bigcap \hat{b}_i}{b_i \bigcup \hat{b}_i} - \frac{B_i \setminus (b_i \bigcap \hat{b}_i)}{B_i})\} \quad (4)$$

B_i: smallest box enclosing b_i and $\widehat{b_i}$

- ✓ Overall loss

$$L = \lambda_{Cls} L_{Cls} + \lambda_{Reg} L_{Reg} + \lambda_{Edges} L_{Edges} + \lambda_{Giou} L_{Giou} \quad (5)$$

Experiment

Dataset:

- ✓ the VerSe 2019 challenge dataset(80:40:40)
- ✓ MICCAI-CSI 2014 challenge
- ✓ an in-house spine dataset

Metrics:

- ✓ L-Error: mean localization error of average Euclidean distance
- ✓ Id-Rate: ratio between the number of correctly identified over total number in a scan

Table 1. Results when evaluated on the VerSe 2019 challenge dataset. R: Refinement

Methods	Test dataset		Hidden dataset	
	Id-Rate (%)	L-Error (mm)	Id-Rate (%)	L-Error (mm)
Christian payer [7]	95.65	**4.27**	94.25	**4.80**
iFLYTEK [7]	96.94	4.43	86.73	7.13
nlessmann [7]	89.86	14.12	90.42	7.04
Spine-Transformers	97.16	8.39	96.06	7.38
Spine-Transformers+R	**97.22**	4.33	**96.74**	5.31

Table 2. Results on the MICCAI-CSI 2014 challenge dataset. R: Refinement

Methods	Average		Cervical		Thoracic		Lumbar	
	Id-Rate	L-Error	Id-Rate	L-Error	Id-Rate	L-Error	Id-Rate	L-Error
Payer [8]	86.1	6.2	–	–	–	–	–	–
Yang [2]	80.0	9.1 ± 7.2	83.0	6.6 ± 3.9	74.0	9.9 ± 7.5	80.0	10.9 ± 9.1
Liao [3]	88.3	6.5 ± 8.6	95.1	**4.5 ± 4.6**	84.0	7.8 ± 10.2	92.2	5.6 ± 7.7
Glocker [1]	70.0	12.4 ± 11.2	80.0	7.0 ± 4.7	62.0	13.8 ± 11.8	75.0	14.3 ± 12.3
Spine-Transformers	91.8	8.6 ± 6.4	97.6	5.5 ± 4.4	88.9	9.2 ± 6.7	91.8	9.8 ± 6.7
Spine-Transformers+R	**92.2**	**4.8 ± 3.6**	**97.6**	5.5 ± 4.4	**89.3**	**5.3 ± 4.2**	**92.7**	**3.7 ± 2.8**

Table 3. Ablation study results.

Architecture	Components				Results	
	Box Detector	InSphere Detector	Edge Loss	Refinement	Id-Rate	L-Error
Spine-Transformers	✓				96.00	8.21 ± 3.71
Spine-Transformers		✓			96.71	8.10 ± 3.58
Spine-Transformers		✓	✓		96.96	8.04 ± 3.63
Spine-Transformers		✓	✓	✓	**98.70**	**3.41 ± 3.09**

Multi-compound Transformer for Accurate Biomedical Image Segmentation

Yuanfeng Ji[1], Ruimao Zhang[2], Huijie Wang[2], Zhen Li[2], Lingyun Wu[3], Shaoting Zhang[3], and Ping Luo[1(✉)]

[1] The University of Hong Kong, Pok Fu Lam, Hong Kong
pluo@cs.hku.hk

[2] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China

[3] SenseTime Research, Beijing, China

Introduction

- Input：2D image

- Task：Semantic Segmentation

- Previous Study:

✓miss learning the cross-scale dependencies of different pixels

✓miss learning the semantic correspondence of different labels

Motivation

TransUNet

✓First, it only uses the self-attention mechanism for context modeling on a single scale but ignores the cross-scale dependency and consistency. The latter usually plays a critical role in the segmentation of lesions with dramatic size changes.

✓Second, beyond the context modeling, how to learn the correlation between different semantic categories and how to ensure the feature consistency of the same category region are still not taken into account.

Contribution

- ✓Propose the MCTrans, which constructs cross-scale contextual dependencies and appropriates semantic relationships for accurate biomedical segmentation.
- ✓A novel learnable proxy embedding is introduced to build category dependencies and enhance feature representation through self-attention and cross-attention, respectively.
- ✓We plug the designed MCTrans into a UNet-like network and evaluate its performance on the six challenging segmentation datasets.

Methods

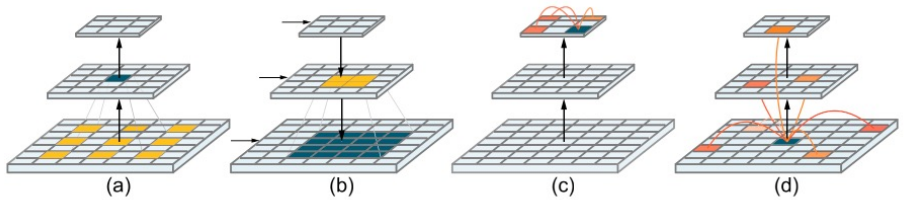


Fig. 1. Conceptual comparison of various mechanisms for context modeling for segmentation. In contrast to (a–c), MCTrans models pixel-wise relationships between multiple scales features, enabling more consistent and effective context encoding. The Prussian blue grids denote the target pixel while other color grids represent the support pixels. For simplicity, we only show a subset of the pathways between target pixels and support pixels. (Color figure online)

- ✓ Use CNN to extract multi-scale features
- ✓ Feed the embedded tokens to TSA module
- ✓ Fold the encoded tokens to several 2D feature maps to generate segmentation

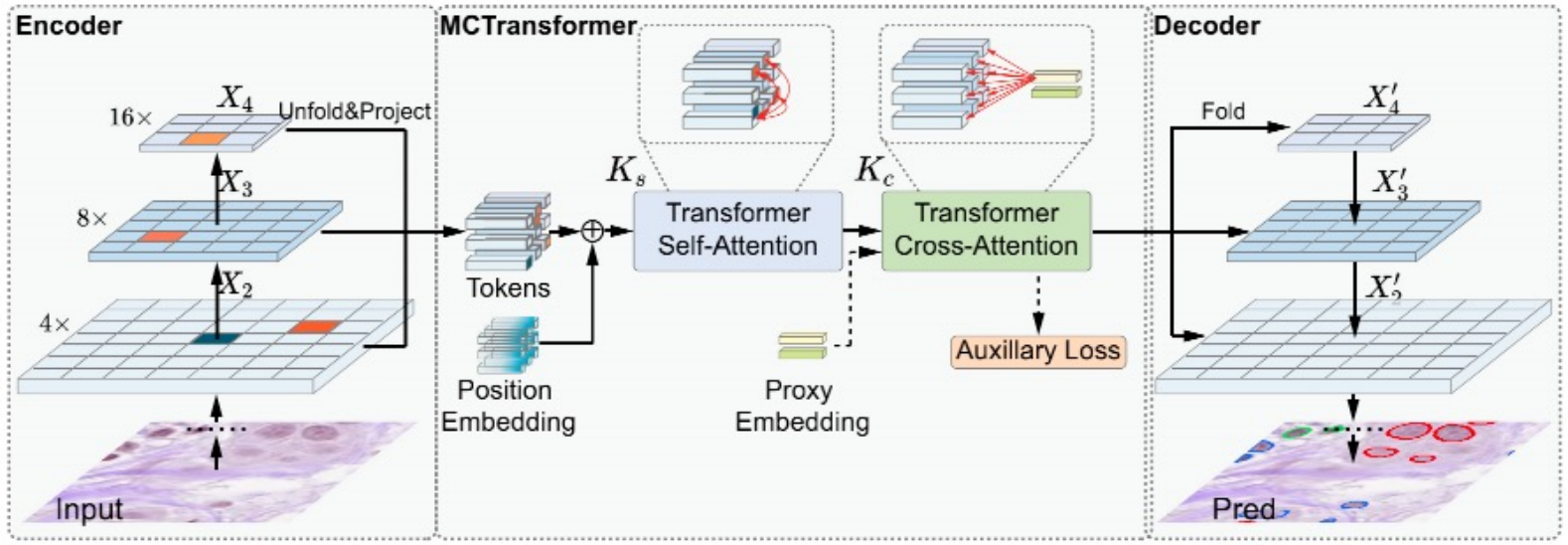


Fig. 2. The overview of MCTrans. We use CNN to extract multi-scale features, and feed the embedded tokens to the Transformer-Self-Attention module to construct the multi-scale context. We add a learnable proxy embedding to learn category dependencies and interact with the feature representations via the Transformer-Cross-Attention module. Finally, we fold the encoded tokens to several 2D feature maps and merge them progressively to generate segmentation results. For the details of the two modules, please refer to Fig. 3.

Methods

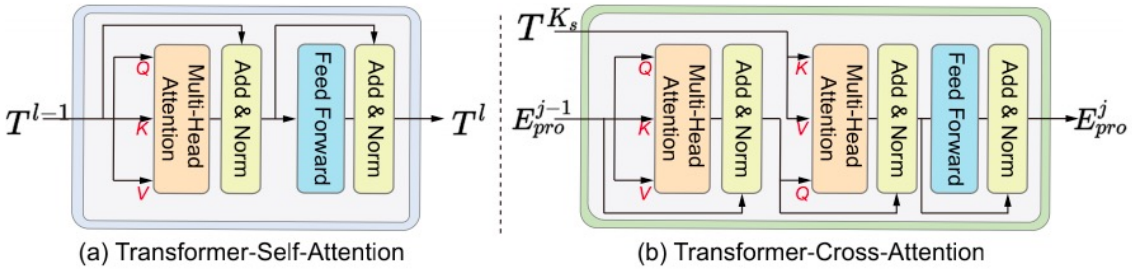


Fig. 3. Illustration of the transformer-self-attention and transformer-cross-attention modules.

✓ TSA module:
encode the contextual information between the multiple features, yielding rich and consistent pixel-level context
✓ TCA module:
introduces learnable embedding for semantic relationship modeling and further enhances feature representations
✓ Patch Size: P*P, P=1
✓ Positional embedding: L*C

$$\left\{X_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times C_i}\right\} \qquad L_i = \frac{HW}{2^{2*i} \times P^2}$$

✓ Concatenate the features of i=2,3,4 level and form overall tokens

$$L = \sum_{i=2}^{4} L_i.$$

✓ Input of self-attention: (q,k,v) T^{l-1}: T+E_{pos} T: token flatten from multi-scale features

$$\text{query} = T^{l-1}\mathbf{W}_Q^l, \text{key} = T^{l-1}\mathbf{W}_K^l, \text{value} = T^{l-1}\mathbf{W}_V^l \qquad (1$$

✓ SA(self-attention)

$$\text{SA}\left(T^{l-1}\right) = T^{l-1} + \text{Softmax}\left(\frac{T^{l-1}\mathbf{W}_Q^l\left(T^{l-1}\mathbf{W}_K^l\right)^\top}{\sqrt{d_k}}\right)\left(T^{l-1}\mathbf{W}_V^l\right) \qquad (2)$$

✓ MSA W_O: parameter of output linear projection

$$\text{MSA}(T^{l-1}) = \text{Concat}\left(\text{SA}_1, \ldots, \text{SA}_\text{h}\right) W_O^l \qquad (3)$$

$$T^l = \text{MSA}\left(T^{l-1}\right) + \text{FFN}\left(\text{MSA}\left(T^{l-1}\right)\right) \in \mathbb{R}^{L \times C} \qquad (4)$$

Methods

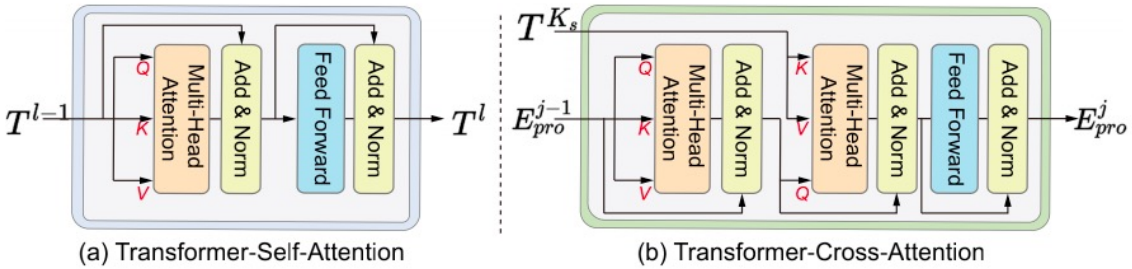
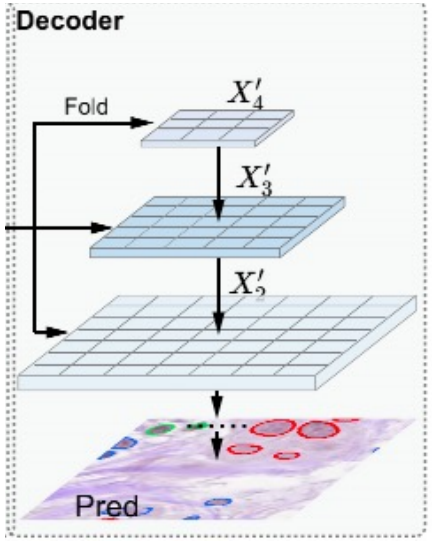


Fig. 3. Illustration of the transformer-self-attention and transformer-cross-attention modules.

- ✓ The output of last layer of TCA is passes to a linear projection head and yields a multi-class prediction
- ✓ Token fold back to 2D features

$$\{X_0, X_1, X_2', X_3', X_4'\}$$

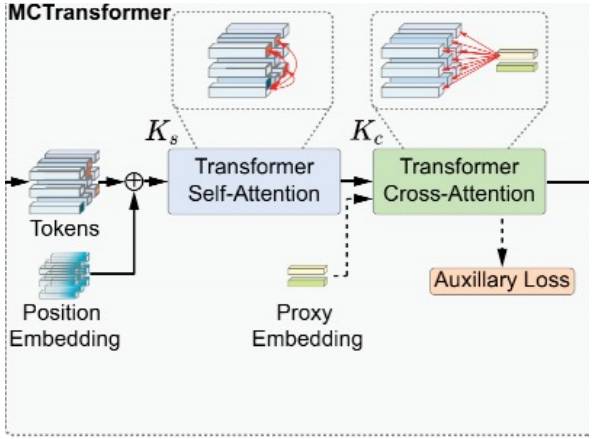


- ✓ Overall token + position embedding->TSA Module->TCA Module+ proxy embedding
- ✓ Proxy embedding: to learn the global semantic relationship between categories
- ✓ M: number of categories of the dataset

$$E_{pro} \in \mathbb{R}^{M \times C}$$

- ✓ E_{pro}: proxy embedding

Yield inputs (q,k,v) of the first MSA block



- ✓ Another MSA block

q: proxy embedding

k,v: token T

- ✓ Deformable Self Attention(Deformable DETR)

Token T has long sequence, computation complexity

Experiment

Table 3. Comparisons with other conventional methods on the Pannuke dataset.

Method	Params (M)	Flops (G)	Neo	Inflam	Conn	Dead	Epi	Ave
UNet [16]	7.853	14.037	82.86	66.16	62.45	38.10	75.02	64.92
UNet++ [24]	9.163	34.661	82.14	66.01	61.61	38.47	76.54	64.97
CENet [9]	17.682	18.779	83.05	66.92	62.41	38.021	76.44	65.37
AttentionUNet [15]	8.382	15.711	81.85	65.37	63.79	38.96	75.45	64.27
MCTrans	7.642	18.065	83.99	68.24	63.95	47.39	78.42	68.40
UNet [16]	24.563	38.257	82.85	65.48	62.29	40.11	75.57	65.26
UNet++ [24]	25.094	84.299	82.03	67.58	62.79	40.79	77.21	66.08
CENet [9]	34.368	41.389	82.73	**68.25**	63.15	41.12	77.27	66.50
AttentionUNet [15]	25.094	40.065	82.74	65.42	62.09	38.60	76.02	64.97
MCTrans	23.787	39.71	**84.22**	68.21	**65.04**	**48.30**	**78.70**	**68.90**

- ✓ Dataset: multi-class segmentation dataset
- ✓ Six dataset(1 for ablation study, 5 for evaluation)

Table 4. Comparisons with other top methods on the five lesion segmentation datasets.

Method	CVC-Clinic	CVC-Colon	ETIS	Kavairs	ISIC2018
UNet [16]	88.59	82.24	80.89	84.32	88.78
UNet++ [24]	89.30	82.86	80.77	84.95	88.85
CENet [9]	91.53	83.11	75.03	84.92	89.53
AttentionUNet [15]	90.57	83.25	79.68	80.25	88.95
MCTrans	**92.30**	**86.58**	**83.69**	**86.20**	**90.35**

- ✓ First Group: VGG-Style network as feature extractor
- ✓ Second Group: ResNet-34
- ✓ Table 4 : on other five lesion segmentation tasks
- ✓ Slight computation with significant improvements

Experiment-Ablation Study

Table 1. Ablation studies of core components of MCTrans. The performance is evaluated on Pannuke dataset. We estimate Flops and parameters by using $[1 \times 3 \times 256 \times 256]$ input. Note that, UNet+VIT-Enc network is equivalent to TransUNet.

Method	Params (M)	GFlops	Neo	Inflam	Conn	Dead	Epi	Ave
UNet [16]	7.853	14.037	82.86	66.16	62.45	38.10	75.02	64.92
UNet [16]+NonLocal [20]	8.379	14.172	82.67	67.48	62.63	40.44	76.41	65.93
UNet [16]+VIT-Enc [7]	27.008	18.936	83.34	68.33	63.18	38.11	77.25	66.04
MCTrans w/o TCA	7.115	18.061	83.87	68.54	64.68	44.25	78.30	67.93
MCTrans w/o TSA	6.167	11.589	83.39	67.82	63.94	44.35	76.31	67.16
MCTrans w/o Aux-Loss	7.642	18.065	83.92	67.92	64.22	45.16	78.14	67.87
MCTrans	7.642	18.065	83.99	68.24	64.95	46.39	78.42	68.40

Table 2. Sensitivity to the number of the TSA and TCA module.

N_s	2	4	6	8	N_c	2	4	6	8
DSC	67.25	67.67	**67.93**	67.50	–	68.15	**68.40**	68.31	68.11

Ablation study:

- ✓ TCA
- ✓ TSA
- ✓ Aux-loss
- ✓ Non-Local:

Experiment-Visualization

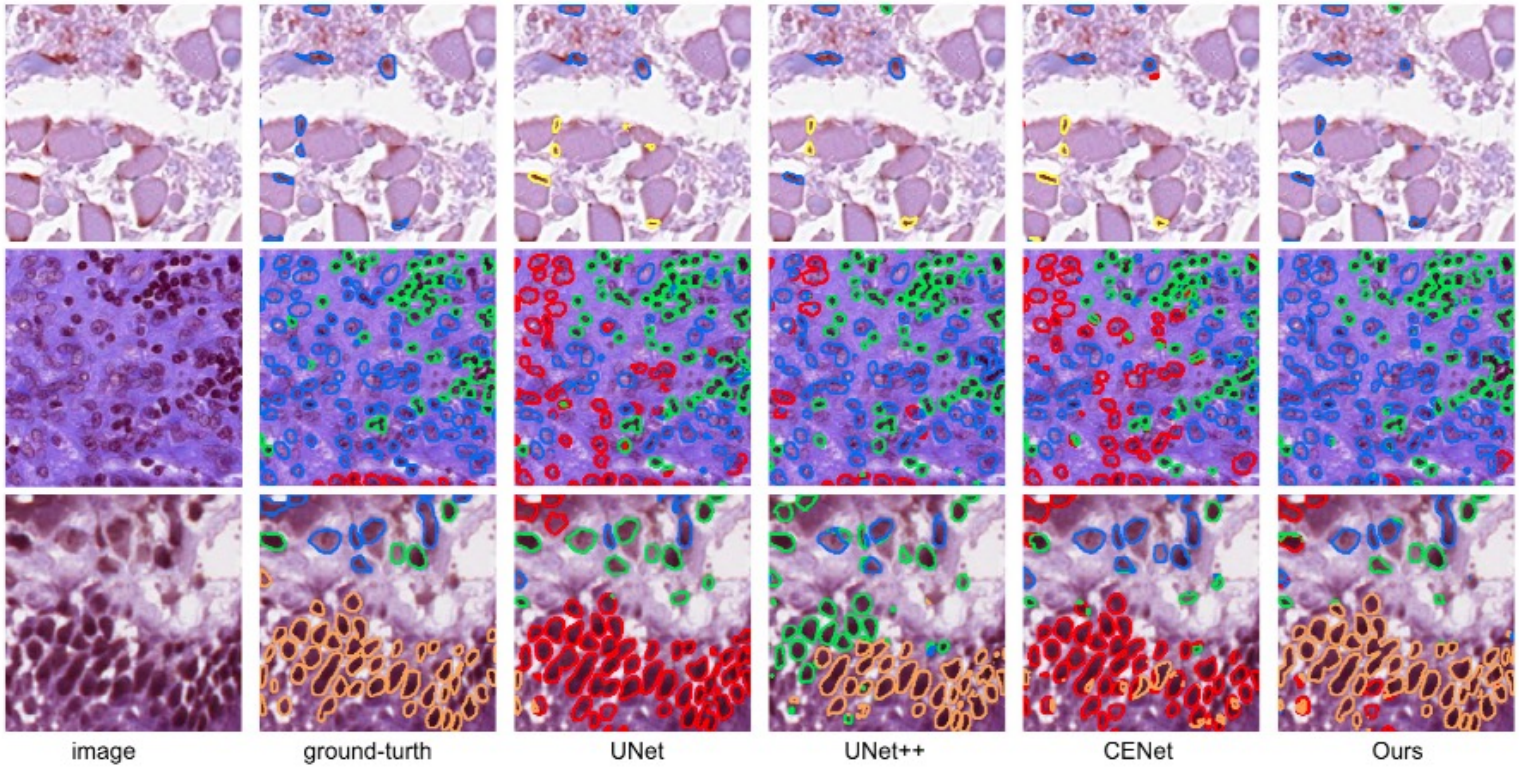


Fig. 4. Segmentation results on the Pannuke dataset, which contains of five foreground classes: Neoplastic, Inflammatory, Connective, Dead, and Non-Neoplastic Epithelial. (Color figure online)

Thank you!