UTNet：基于UNet和Transformer架构的医学图像分割探索

展示成员：鹿逸文

目录

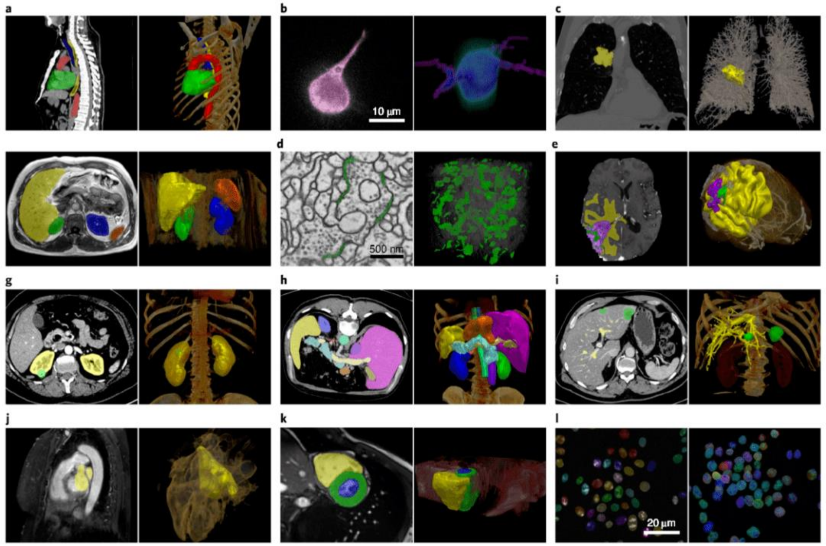
1. 项目背景与意义

项目背景

- 医学图像分割是医学领域中至关重要的任务，它在诊断、治疗规划和疾病监测等方面发挥着关键作用

- 传统的医学图像分割受限于主观性，且人力成本较高

- Transformer在自然语言处理领域取得了显著成就，但在医学视觉领域的应用尚未深入探索。



意义

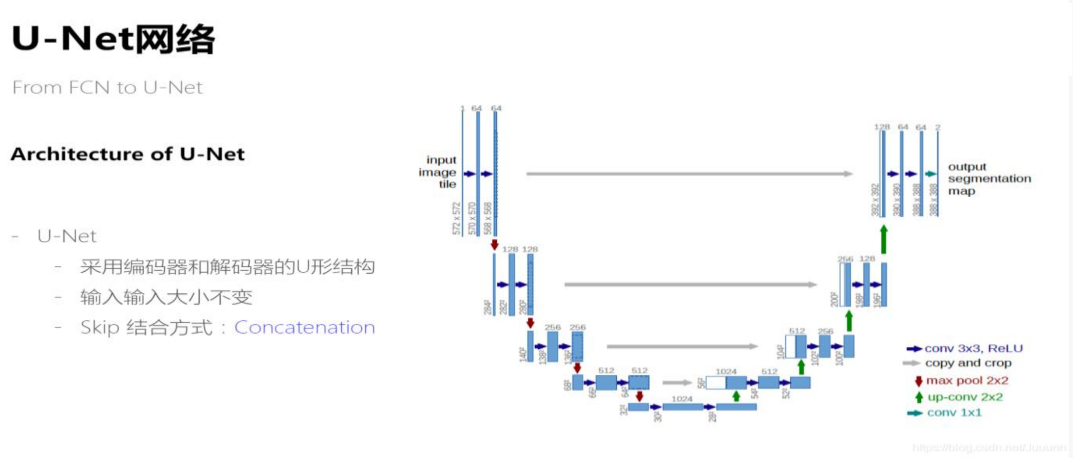
UTNet有望为疾病的早期检测、治疗规划和病情预测提供有力的支持，提高医疗诊断的精确性和效率，最终改善患者的治疗结果和生存率。

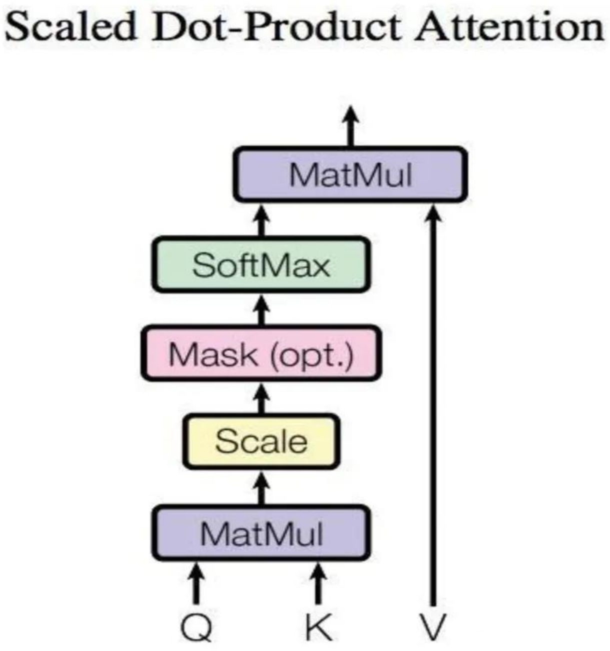
2. 主要贡献

整合了卷积和注意力策略的优势

提出了一种高效的自注意机制

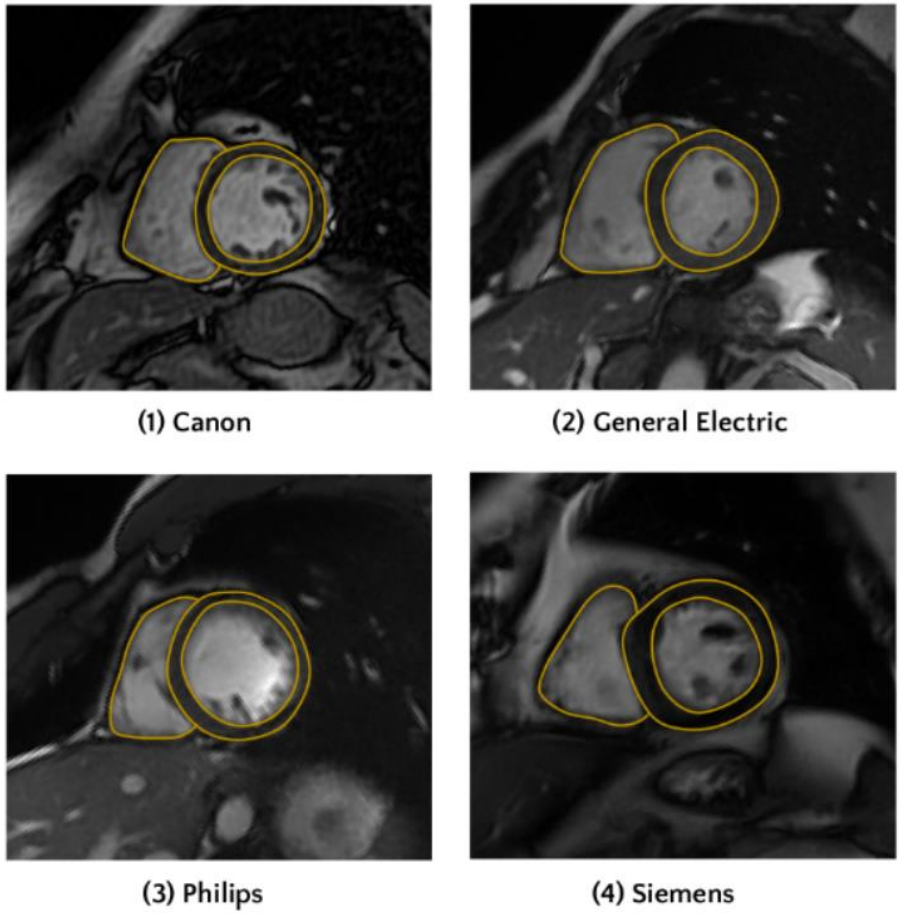
引入了全新的自注意解码器





3. 数据处理

数据集选择 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge



所用的MRI图像是心脏电影成像序列，是带时间维度的三维心脏图像。

来源

数据集来源于多中心，多供应商和多疾病心脏图像分割挑战赛。由375名肥厚性和扩张型心肌病患者以及健康受试者组成。

标签

训练集包含150个带标注的图像。由来自各个机构经验丰富的临床医生对CMR图像进行了分割，包括左（LV）和右心室（RV）血池以及左心室心肌（MYO）的轮廓。标签为：0（背景），1（LV），2（MYO）和3（RV）。

3. 数据处理

数据预处理

01 缩放

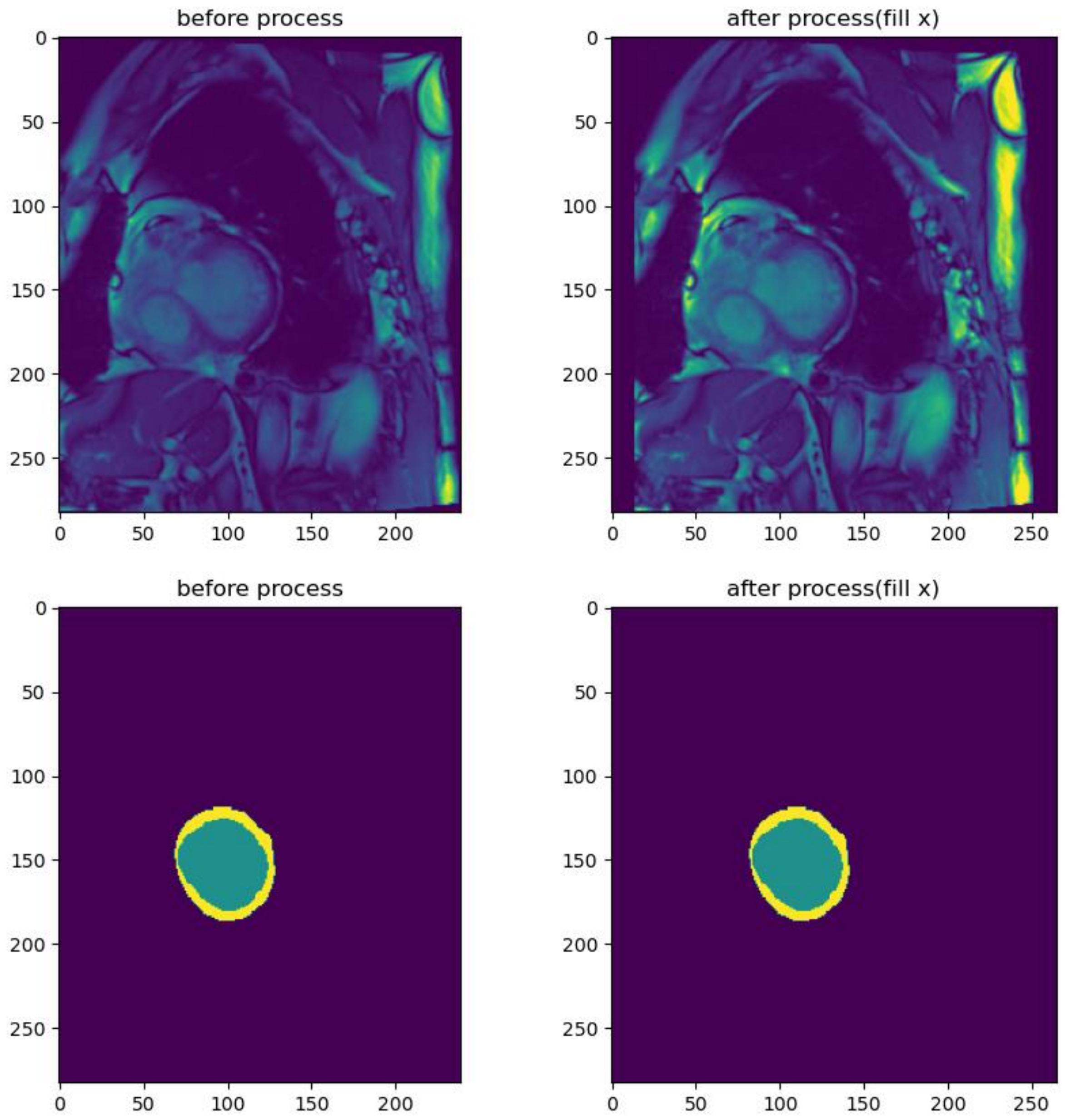
- 将所有数据重新采样为 x-y 平面中 1.2x1.2 mm 的间距

02 截断

- 对缩放后的原始数据进行异常值截断处理，将98%以上的灰度值和5%以下的灰度值进行截断

03 归一化

- 采用均值为0，方差为1的方式对原始图像进行归一化处理



3. 数据处理

数据增强

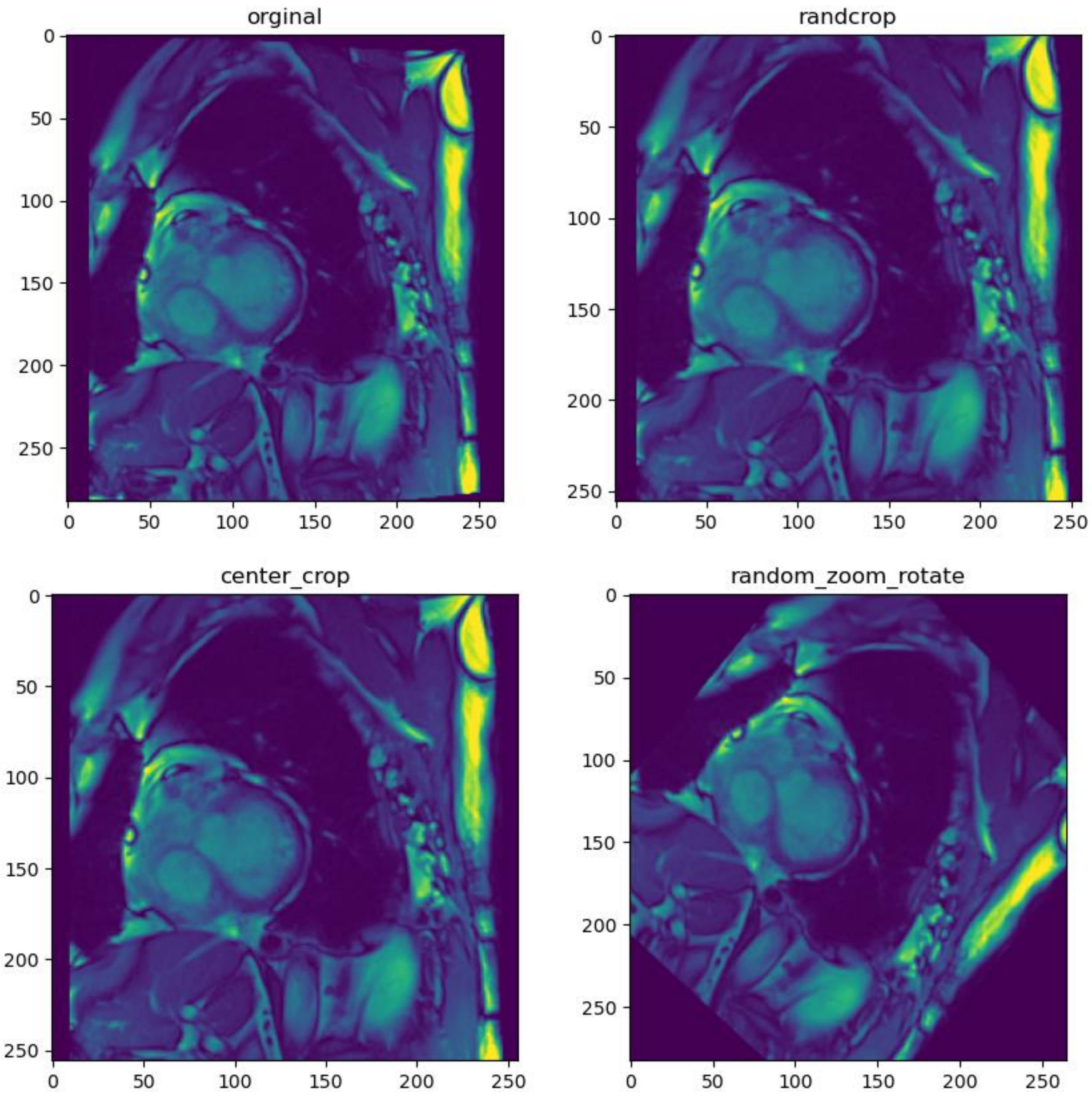
数据增强

1 添加高斯噪声

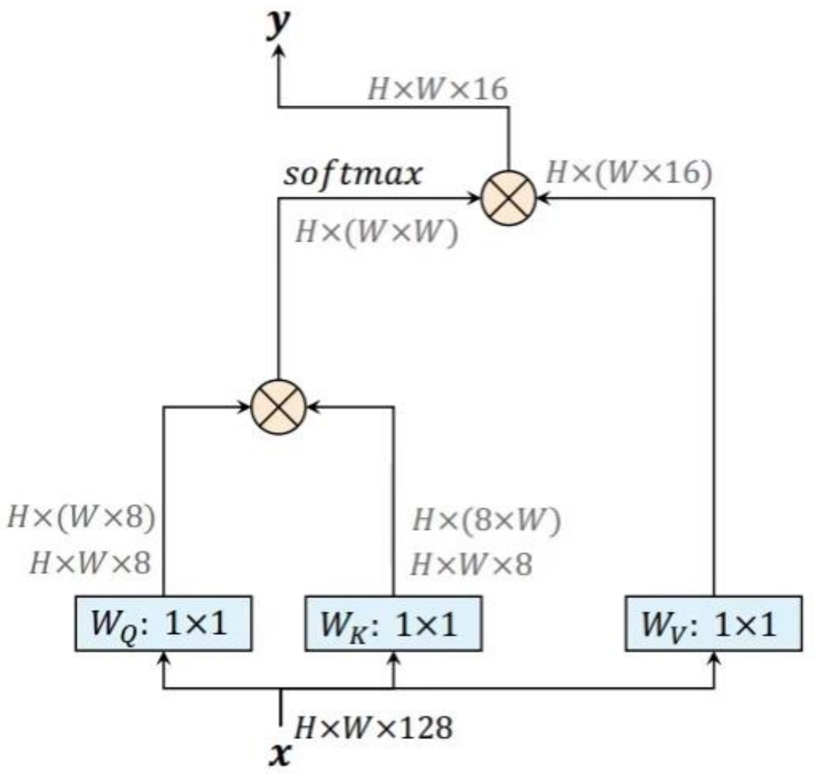
2 亮度随机调整

3 Gamme校正对比度

4 随机缩放、旋转、裁剪

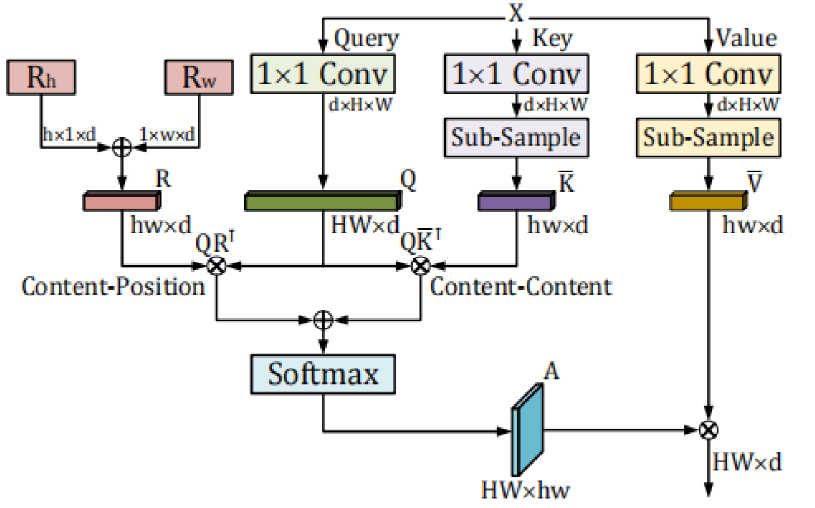


4. 模型构建：高效自注意力机制

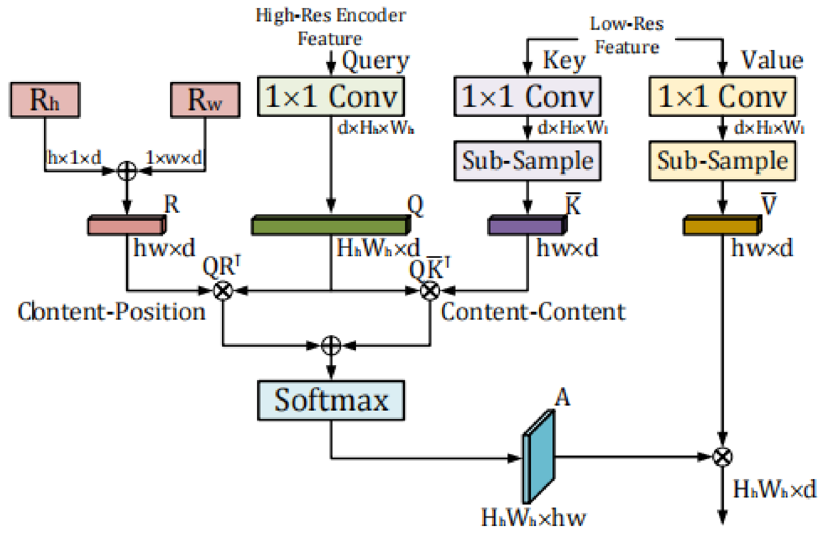


原生注意力机制

$$\mathrm{Attention}(\mathbf{Q},\overline{\mathbf{K}},\overline{\mathbf{V}}) = \underbrace{\mathrm{softmax}\left(\frac{\mathbf{Q}\overline{\mathbf{K}}^{\mathsf{T}}}{\sqrt{d}}\right)}_{\overline{P}:n\times k}\underbrace{\overline{\mathbf{V}}}_{k\times d}$$



多头注意力编码器



多头注意力解码器

4. 模型构建：相对位置编码

0.3	0.9	0.4	0.5
0.2	0.3	0.2	0.4
0.7	0.1	0.3	0.9
0.6	0.7	0.2	0.3

相对位置偏置矩阵

4	3	1	0
5	4	2	1
7	6	4	3
8	7	5	4

相对位置索引矩阵

数组维度

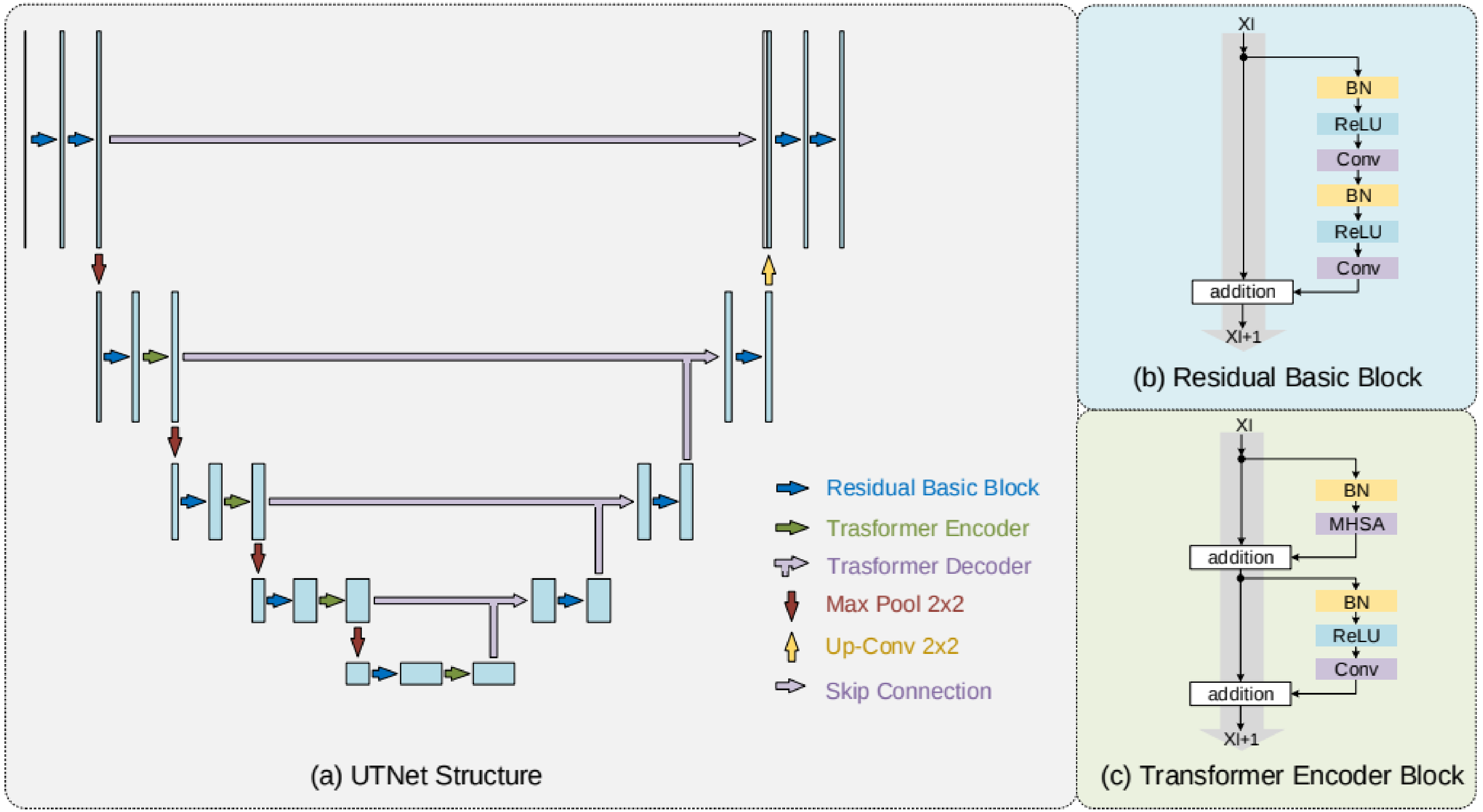
可学习数组

相对位置偏置表 (2M-1)x(2M-1)

0.5	0.4	0.2	0.9	0.3	0.2	0.1	0.7	0.6
0	1	2	3	4	5	6	7	8

$$\mathrm{Attention}(\mathbf{Q}, \overline{\mathbf{K}}, \overline{\mathbf{V}}) = \underbrace{\mathrm{softmax}\left(\frac{\mathbf{Q}\overline{\mathbf{K}}^{\mathsf{T}} + \mathbf{S}_H^{rel} + \mathbf{S}_W^{rel}}{\sqrt{d}}\right)}_{\overline{P}: n \times k} \underbrace{\overline{\mathbf{V}}}_{k \times d}$$

4. 模型构建：UNet+Attention



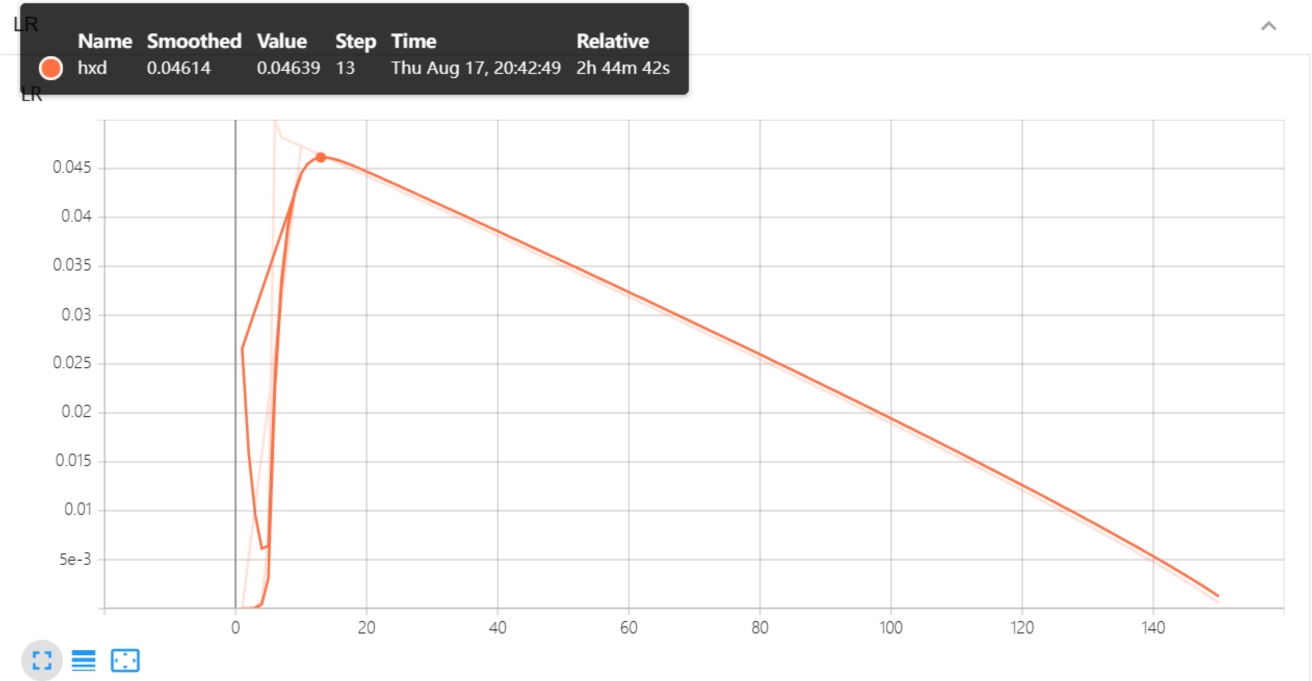
UTNet架构

5. 训练技巧:学习率调度器

1 学习率的大小很重要

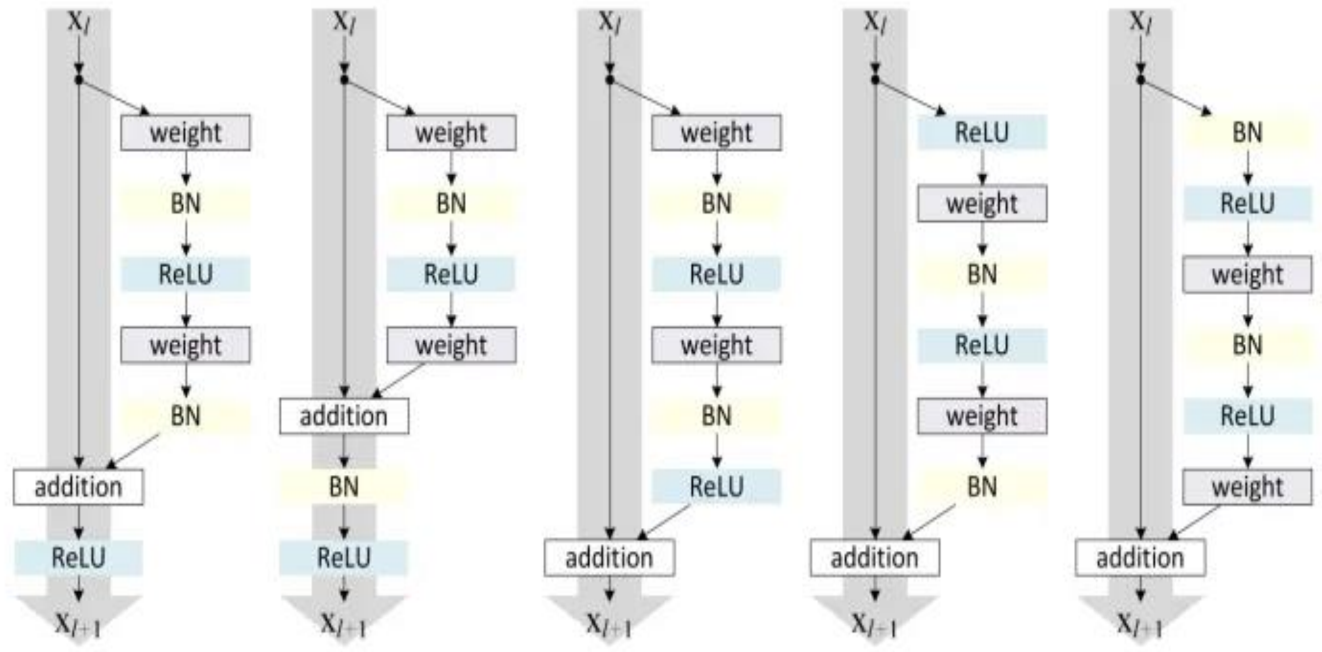
2 衰减速率同样很重要

3 预热阶段

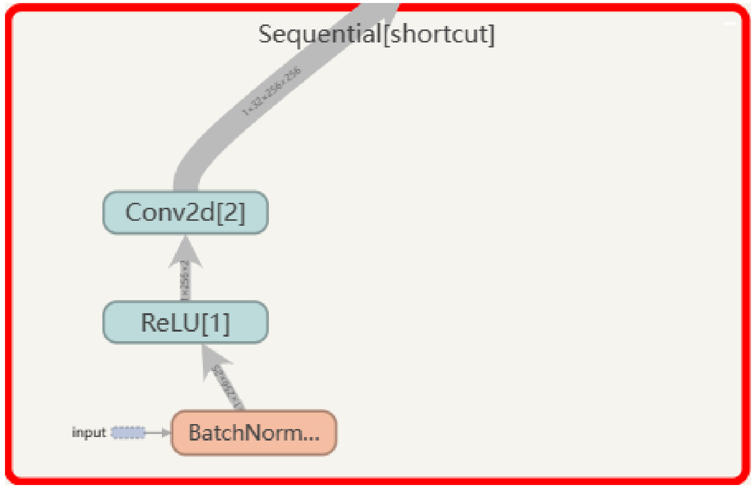


训练过程中**学习率**变化

5. 训练技巧:pre-activation

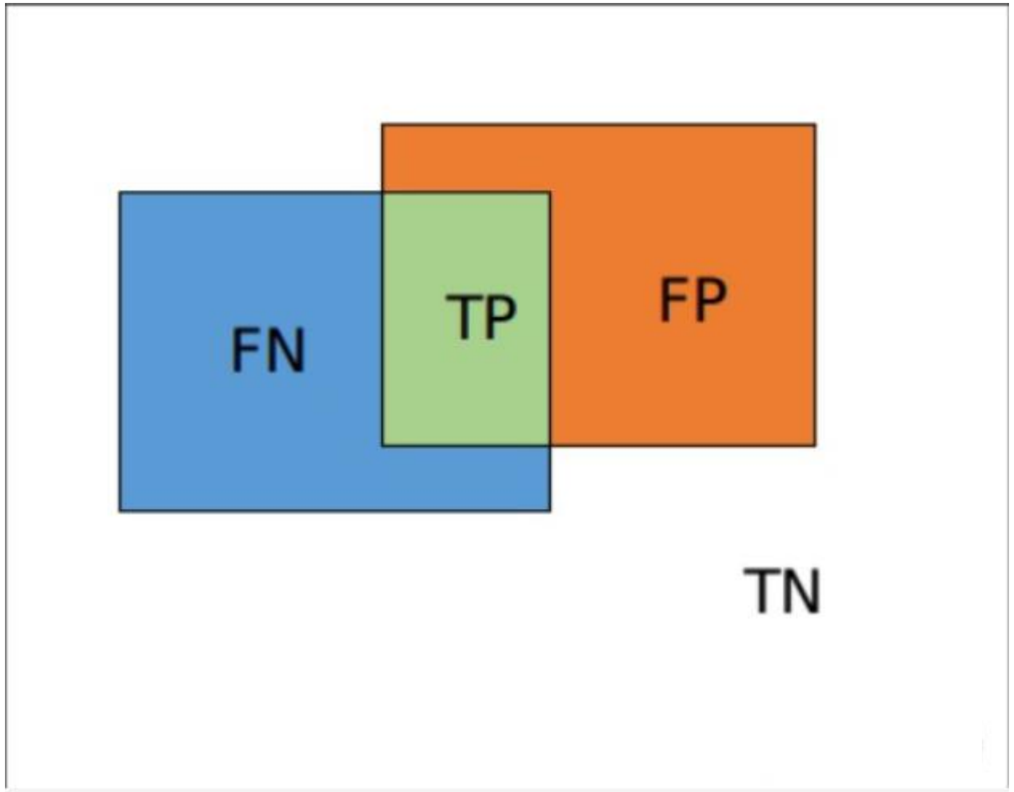


(a) original (b) BN after addition (c) ReLU before addition (d) ReLU-only pre-activation (e) **full pre-activation**



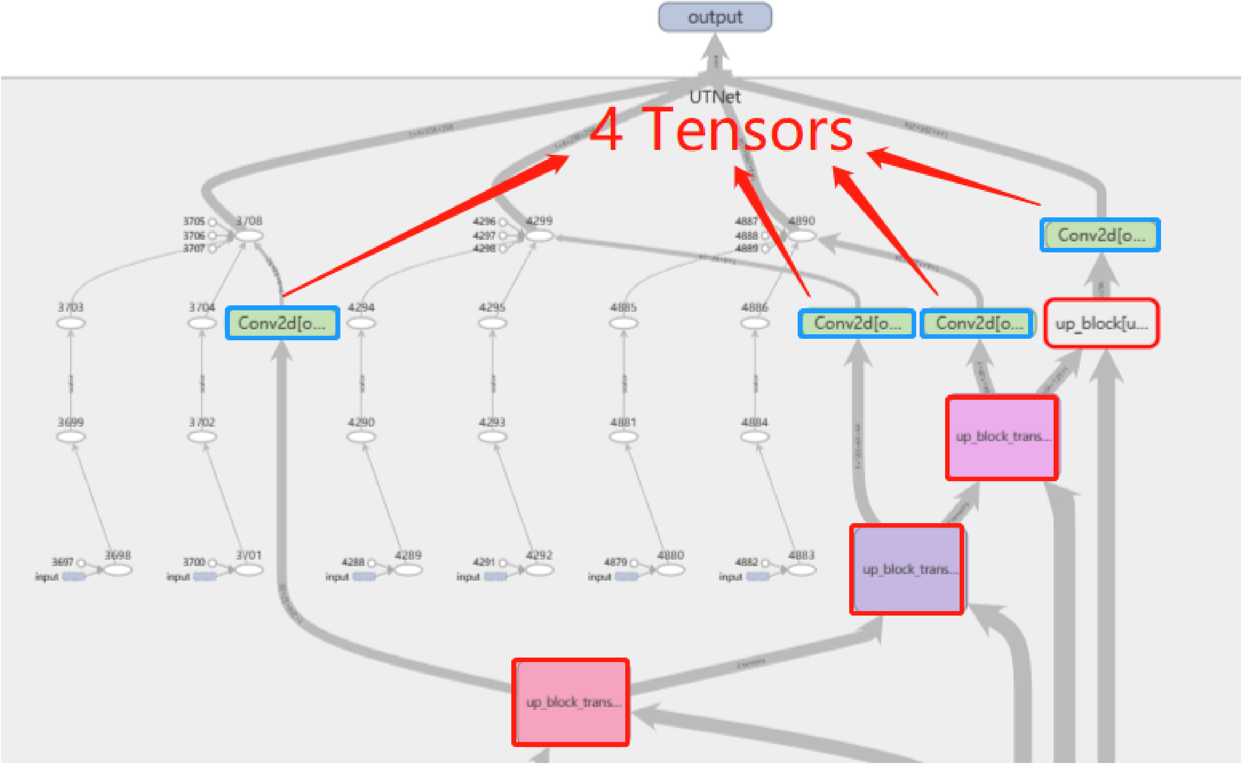
UTNet中的残差部分

5. 训练技巧:损失函数



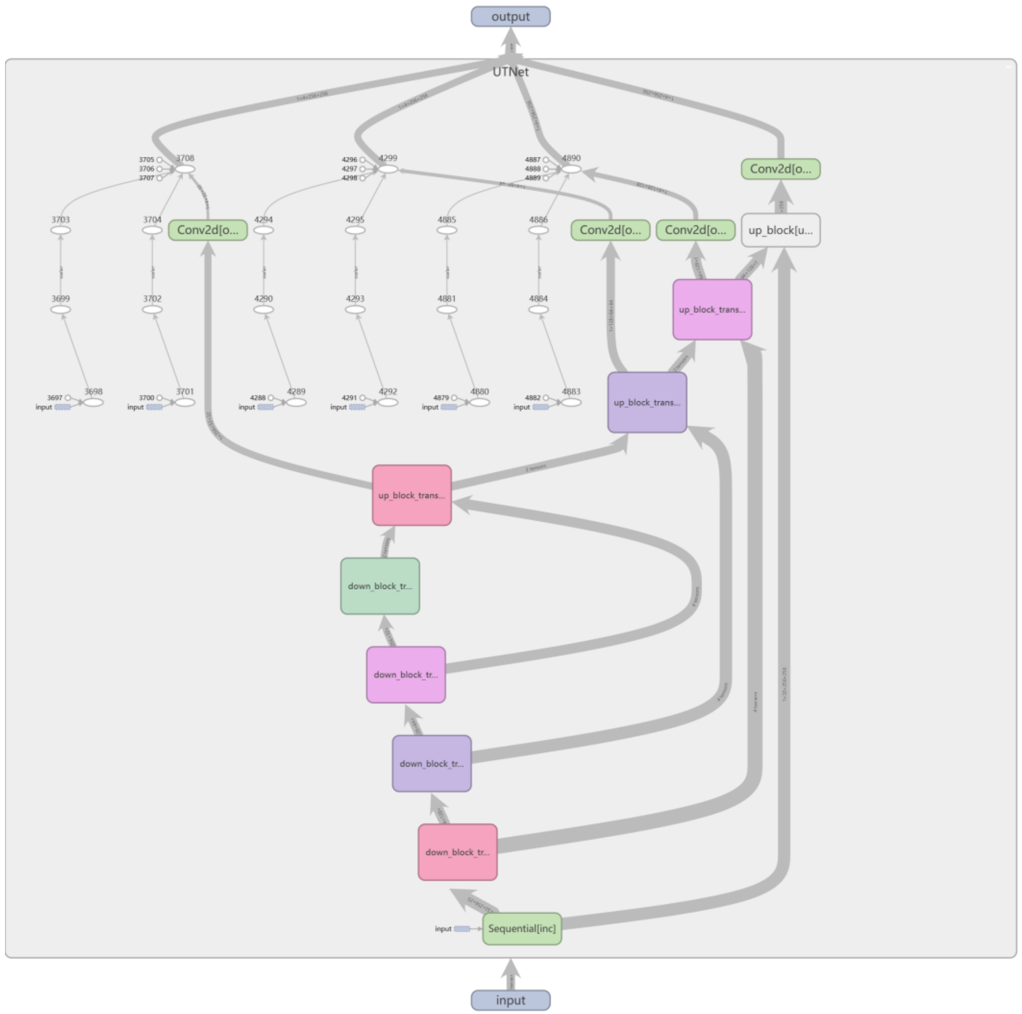
$$dice = \frac{2TP}{2TP+FP+FN}$$

$$DiceLoss = 1 - Dice$$

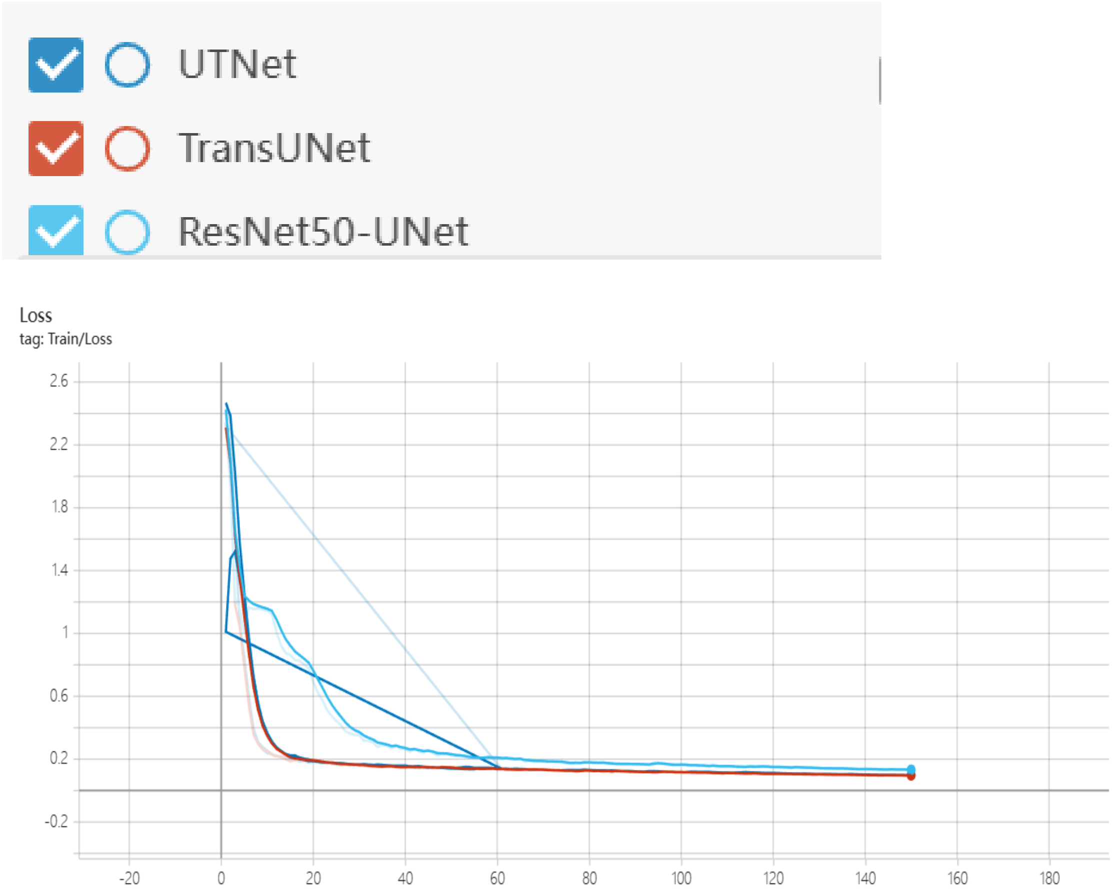


Aux_loss

6. 实验：模型训练

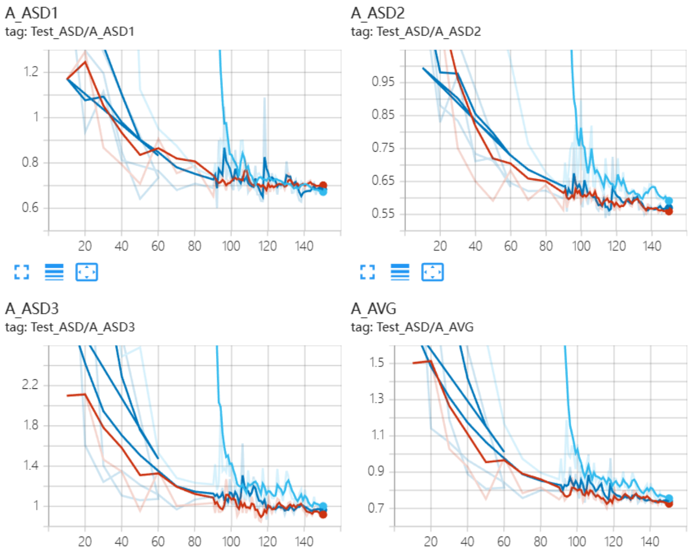


UTNet的TensorBoard可视化

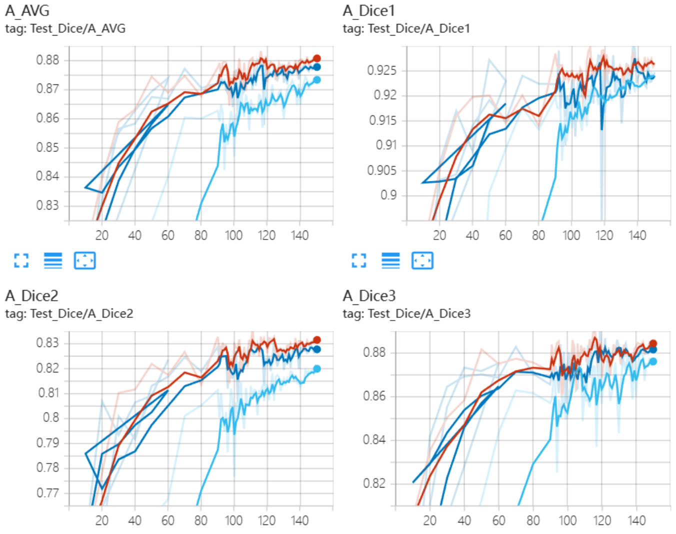


Train_loss

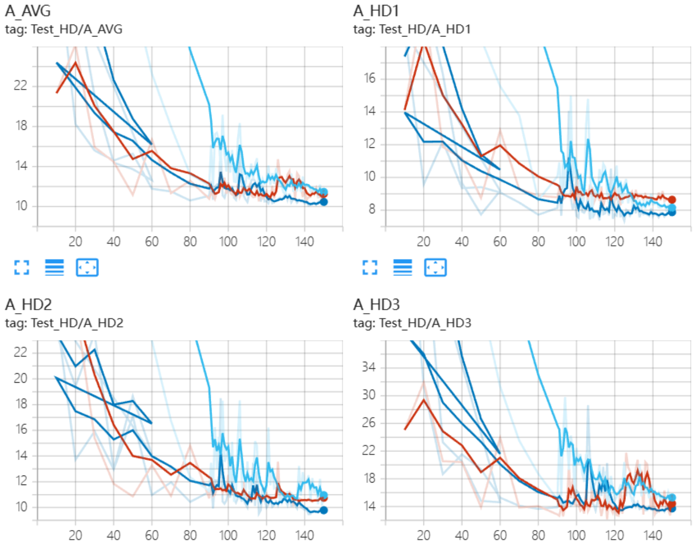
6. 实验：模型测试—指标评估



平均表面距离



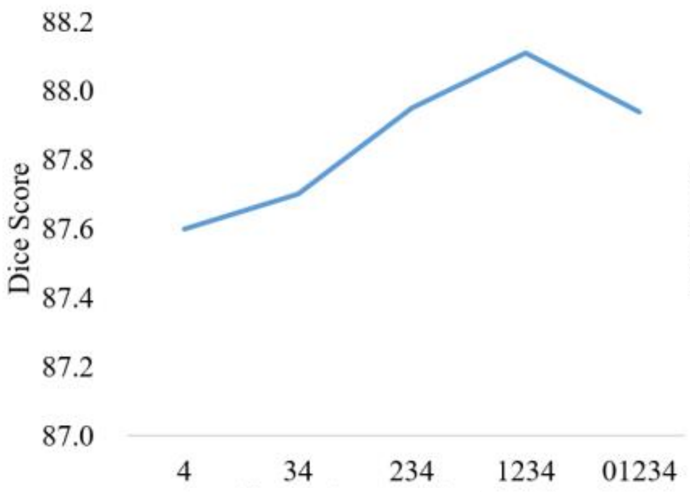
Dice系数



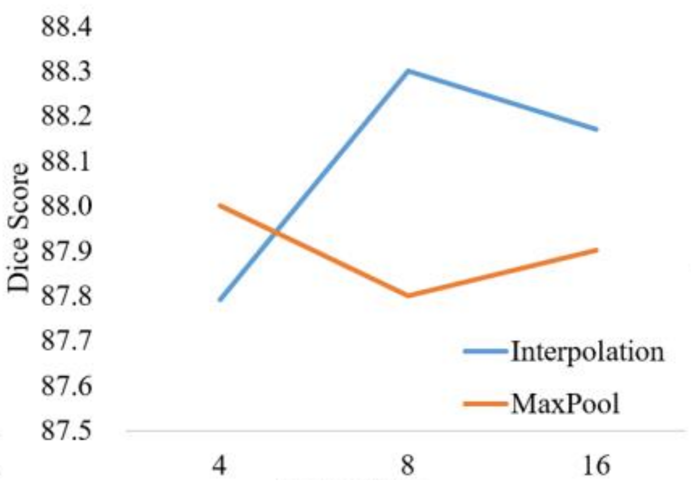
鲁棒豪斯多夫距离

6. 实验：模型测试---消融实验

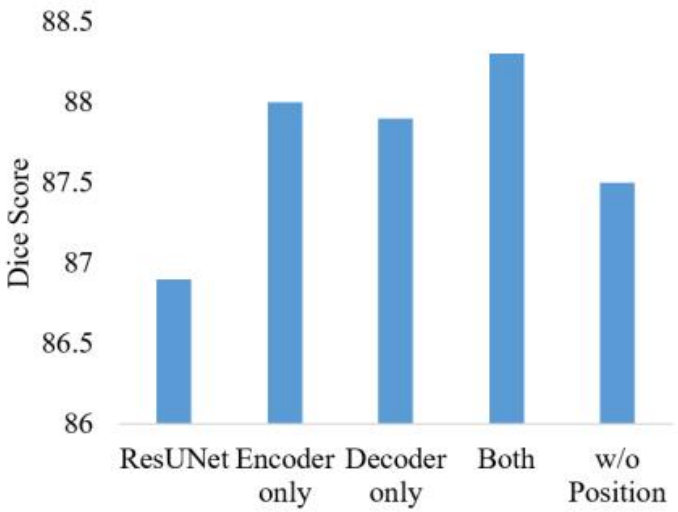
消融实验：



注意力机制放置位置



缩小尺寸

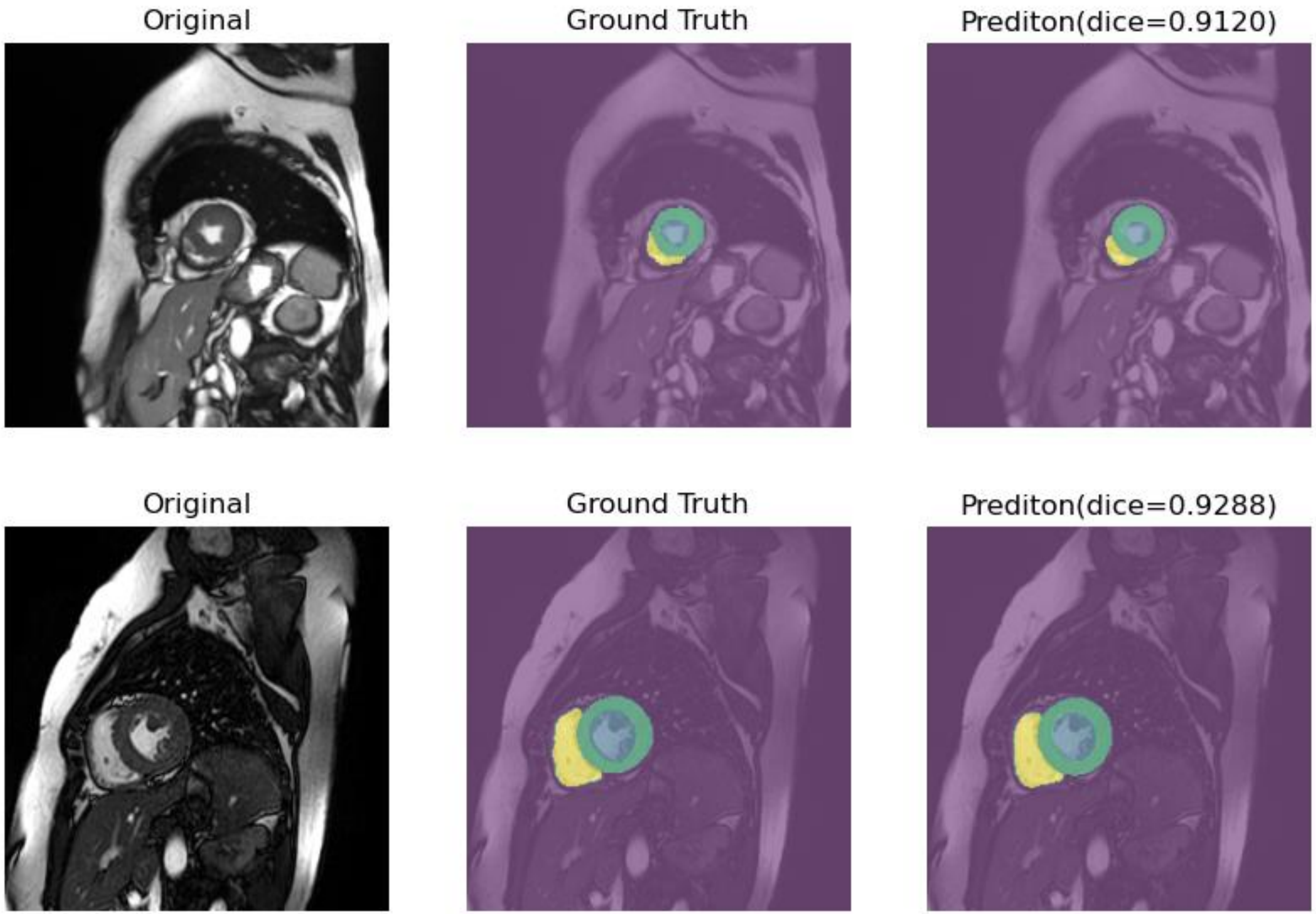


模块组合

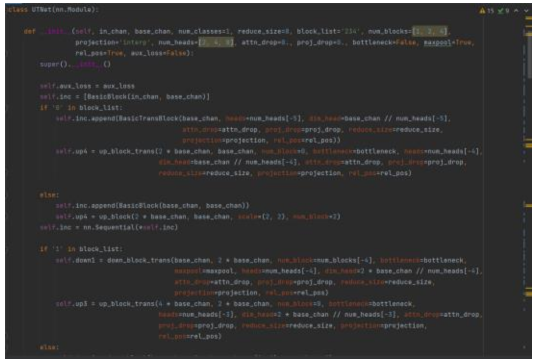
6. 实验：模型效果

供应商C

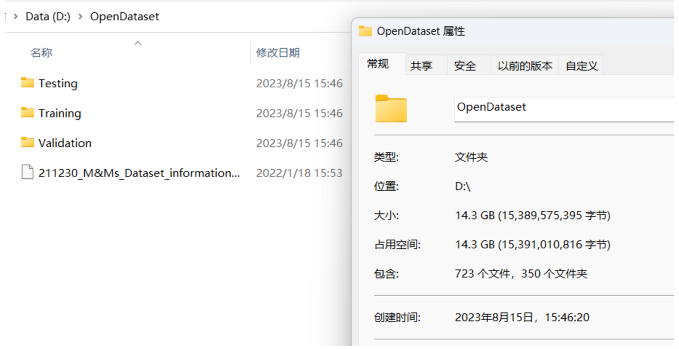
供应商D



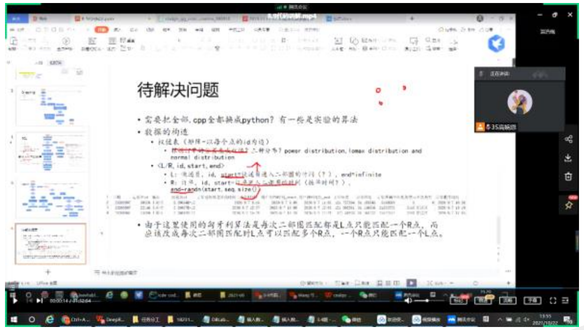
总结



建立模型，编写
代码1000+



处理数据
14G+



开展会议讨
论10余次

阅读文献
5+篇

UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation

Yunhe Gao[1], Mu Zhou[1,2], and Dimitris Metaxas[1]

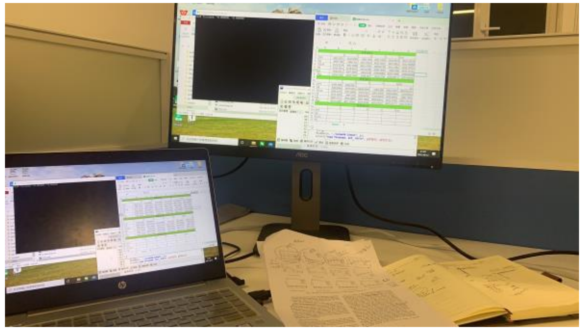
[1] Department of Computer Science, Rutgers University
[2] SenseBrain and Shanghai AI Laboratory and Centre for Perceptual and Interactive Intelligence

Abstract. Transformer architecture has emerged to be successful in a number of natural language processing tasks. However, its applications to medical vision remain largely unexplored. In this study, we present UTNet, a simple yet powerful hybrid Transformer architecture that integrates self-attention into a convolutional neural network for enhancing medical image segmentation. UTNet applies self-attention modules in both encoder and decoder for capturing long-range dependency at different scales with minimal overhead. To this end, we propose an efficient self-attention mechanism along with relative position encoding that reduces the complexity of self-attention operation significantly from $O(n^2)$ to approximate $O(n)$. A new self-attention decoder is also proposed to recover fine-grained details from the skipped connections in the encoder. Our approach addresses the dilemma that Transformer requires huge amounts of data to learn vision inductive bias. Our hybrid layer design allows the initialization of Transformer into convolutional networks without a need of pre-training. We have evaluated UTNet on the multi-label, multi-vendor cardiac magnetic resonance imaging cohort. UTNet demonstrates superior segmentation performance and robustness against the state-of-the-art approaches, holding the promise to generalize well on other medical image segmentations. Code is available[3].

1 Introduction

Convolutional networks have revolutionized the computer vision field with outstanding feature representation capability. Currently, the convolutional encoder-decoder architectures have made substantial progress in position-sensitive tasks, like semantic segmentation [14,11,20,17,6]. The used convolutional operation captures texture features by gathering local information from neighborhood pixels. To aggregate the local filter responses globally, these models stack multiple

20h+实验，
修改代码



请老师批评指正！