暨南大學
JINAN UNIVERSITY

本科课程论文

（2021 — 2022 学年第 一 学期）

论文题目：基于 MONAI 框架实现食管癌 GTV 分割

课 程 名 称：机器学习

课 程 类 别：专业必修课

学 生 姓 名：伍光恒

学　　号：2019051306

学　　院：信息科学技术学院

专　　业：智能科学与技术

任 课 教 师：段俊伟

教 师 单 位：信息科学技术学院

2021 年 12 月 14 日

基于 MONAI 框架实现食管癌 GTV 分割

伍光恒
信息科学技术学院
暨南大学
广州，中国
wuguangheng2002@gmail.com

摘要—肿瘤总体积（GTV）的划定是癌症放射治疗计划中的一个关键步骤。GTV 描述了肿瘤大体上的主要治疗区域，而自动 GTV 分割非常具有挑战性：GTV 分割依赖于放射治疗计算机断层扫描（RTCT）图像的肿瘤外观，但肿瘤与周围组织的对比度较低，还存在食道发生变异以及偶尔存在异物等情况。目前在医学图像处理方面并无比较统一的标准，而 MONAI 是目前一个比较成熟的用于医疗保健领域的框架，它简化了医学图像处理的流程，使得更多医学相关工作者能更快构建适合的模型。鉴于此，本文基于 MONAI 构建了两个模型用于食管癌 GTV 分割任务：3D Unet 与 UNETR。实验结果表明：使用 MONAI 框架能够快速构建所需模型，但在本次实验中使用 3D U-Net 与 UNETR 完成食管癌 GTV 分割并未取得令人满意的结果，还有待调整和改进。

***Index Terms*—食管癌，MONAI 框架，GTV（Gross Tumor Volume），分割，3D Unet，UNETR**

I. Introduction

食道癌的死亡率在全世界所有癌症中排名第六，占癌症死亡人数的二十分之一 [1]。因为它通常在晚期阶段才被诊断出来。放射疗法（RT）通常是主要的治疗方式之一。在 RT 治疗中最关键和最具有挑战性的任务是肿瘤总体积 (GTV) 和临床靶体积 (CTV) 的勾画。GTV 表示可见的肿瘤区域，CTV 描述了覆盖微观肿瘤区域的区域。目前的临床方案主要依赖于手动 GTV 和 CTV 勾画，这既费时又费力，并且不同医师勾画的区域存在不完全一致的可能。因此而寻求 GTV 和 CTV 分割的自动化方法，这可能会提高目标轮廓的准确性和一致性，并且能显著缩短勾画时间。目前也有许多基于机器学习或者深度学习提的出肿瘤 3D 分割方法，比如 [2] 提出了 PSNN 模型，[3] 提出的 DDAUnet 等。但这些方法受限于各自不同的结构，并不具备即插即用，个性化定制结构的优势，需要较多训练技巧才能达到较好的效果。另外，据调研，在食道癌的 3D 分割任务上，目前仍不能实现较好的分割效果。

II. Related Works

一些研究表明已经可以解决自动食道 GTV 分割的问题 [4]。[4] 开发了一种具有最大曲率策略的自适应区域生长算法，以单独将食管肿瘤分割。但是，由于 PET 和 RTCT 之间的错位以及它们的不同的成像原理，即使是专用的跨模式注册算法也可能无法实现令人满意的结果。近年来，深度学习在医学图像分析领域引发了广泛的关注。然而，深度学习技术很少用于食道分割，甚至更少用于食管肿瘤分割任务。[5] 提出了一种用于 3D CT 上的食道分割的全卷积神经网络（FCNN），在该研究中，50 个 CT 样本被用作训练集，20 个 CT 样本作为测试集，测试集的平均 DSC 值为 0.76。紧接着 Trullo 等人提出了用于 2D 食道分割的半自动两阶段 FCNN [6]，在第一阶段进行多器官分割，以提取包含食道的 ROI。然后，手动裁剪的 ROI 被送入第二个网络完成食道分割，在测试集上取得了 0.72 的 DSC 值，其中 25 个 CT 图像作为训练集和 30 个 CT 图像作为测试集。[7] 提出了一种基于空间上下文编码的食管临床目标体积（CTV）深度分割框架，以产生基于卓越的基于边缘的 CTV 边界。另一个工作 [8] 提出了一种双流链 3D CNN 融合方法，使用 CT 和 PET 扫描图像来进行分割食道 GTV。他们通过对 110 名患者的扫描图像进行 5 折的交叉验证来评估它们的方法，他们使用 PET 图像作为互补信息可以将 DSC 值从 0.73 增加到 0.76。总的来说，虽然可以在前面提到的方法中获得合理的结果，但是在没有额外的知识限制的情况下，CT 模型中食管 GTV 分割的问题仍被称为病态问题，并且

仍然具有非常大的挑战性。上面提及的大多数工作解决了食道分割的问题，而不是食道肿瘤分割。然而，由于肿瘤与其相邻组织的对比度较差，使食管肿瘤分割成为一种更复杂的任务。

III. Methods

A. 3D U-Net

3D U-Net 采用与 2D U-Net（U-Net）相同的网络结构，但将所有的 2D 操作，替换为相应的 3D 操作。图 1展示了 3D Unet 的网络架构。与标准 U-net 一样，它有一个分析路径和一个合成路径，每个路径有 4 个分辨率步长。在分析路径，每层包含两个 $3 \times 3 \times 3$ 卷积，每个卷积操作后是一个线性整流单元（Relu），然后是 $2 \times 2 \times 2$ 的最大池化层，其中每个维度的步长为 2。在合成路径中，每层首先是 $2 \times 2 \times 2$ 上卷积操作，每个维度中步长为 2，其次是两个 $3 \times 3 \times 3$ 卷积层，紧接着是一个线性整流单元（Relu）。分析路径到合成路径同一分辨率层之间的连接为合成路径提供了重要的高分辨率特征。在最后一层中，$1 \times 1 \times 1\times$ 卷积减少了输出通道的数量，为了避免了瓶颈，通过将在最大池化之前的通道数加倍。在合成路径中也采用了相同的策略。

另外，3D Unet 还在每个 Relu 之前引入批量归一化操作（BN）。每批在训练期间使用其平均值和标准偏差进行归一化，同时使用这些值更新全局统计信息。这个操作放在具有明确学习尺度和偏好的网络层之后。在测试的时候，通过这些已计算的全局统计数据，学习尺度和偏好来完成归一化。但是，在数据批次大小较小的应用中，使用局部统计数据也在测试时候表现得比较好。

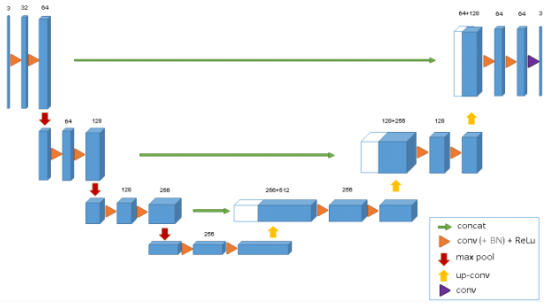


Fig. 2: The 3D u-net architecture. Blue boxes represent feature maps. The number of channels is denoted above each feature map.

图 1. Architecture of 3D UNet [9]

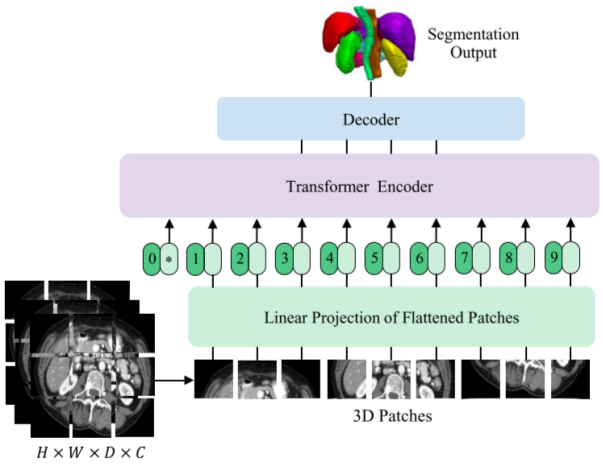


Figure 1. Overview of UNETR. Our proposed model consists of a transformer encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connection.

图 2. Overview of UNETR [10]

B. UNETR

受 Transformer 在自然语言处理领域巨大成功的启发，Ali Hatamizadeh 等人将 3D 医学图像分割重整为为序列到序列预测问题，提出了一个新颖的基于 Transformer 实现的 3D 医学图像分割架构。像 Unet 的结构一样，UNETR 也由收缩路径与扩张路径组成，但其中的编码器是由许多 Transformer 堆叠而成的，编码器通过跳跃连接连接到解码器。在 NLP 应用中，Transformer 通常处理 1D 的输入嵌入序列。UNETR 也采用了相同的方式。UNETR 从 3D 输入 $x \in R^{H \times W \times D \times C}$ 中创建 1D 序列。这里 (H, W, D) 为分辨率，C 为输入的通道数。UNETR 将 x 分割成不重叠的图像块 $x_v \in R^{N \times (P^3 \cdot C)}$，这里 (P, P, P) 表示图像块的分辨率，$N = \frac{(H \times W \times D)}{P^3}$ 为 1D 序列的长度。之后的自注意力计算操作基本上与 ViT（Vision Transformer）相似，简单描述为：得到的 1D 序列经过线性层投影至 K 维的嵌入空间，另外给每个系列佳航一个可学习的位置嵌入编码（采用相加方式），UNETR 完成的是分割任务，不需要像 ViT 一样另外加入分类嵌入向量，然后经过多头注意力计算（MSA）层和多层感知机（MLP）。值得一提的是，UNETR 所采用的解码器是基于 CNN 实现的，这是由于 Transformer 并不能很好的捕捉局部信息，尽管它能很好的捕捉全局信息。UNETR 整体的结构如图 2所示。另外，UNETR 采用了与 Unet 相似的结构，这里不详细展开说明，具体的结构如图 3所

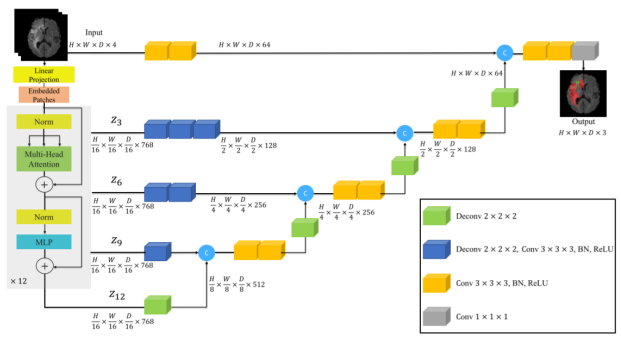
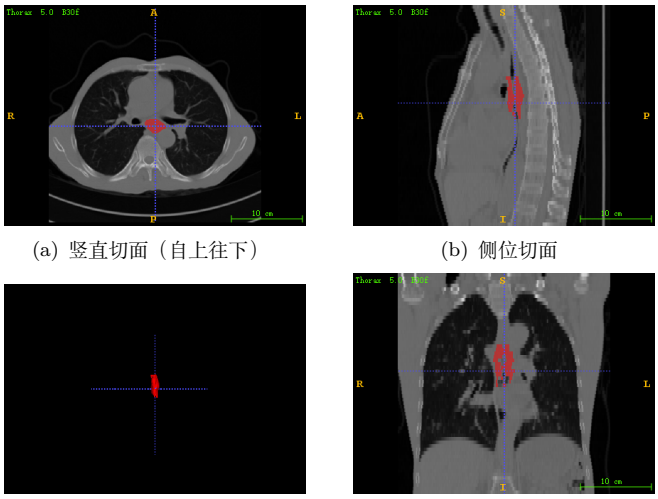


Figure 2. Overview of UNETR architecture. A 3D input volume (e.g. $C = 4$ channels for MRI images), is divided into a sequence of uniform non-overlapping patches and projected into an embedding space using a linear layer. The sequence is added with a position embedding and used as an input to a transformer model. The encoded representations of different layers in the transformer are extracted and merged with a decoder via skip connections to predict the final segmentation. Output sizes are given for patch resolution $P = 16$ and embedding size $K = 768$.

图 3. Overview of UNETR architecture [10]



(a) 竖直切面（自上往下）　(b) 侧位切面

(c) 肿瘤区域（VOI）　(d) 正位切面

图 4. 样本图像示例

示，实现细节可参考原论文 [10]。UNETR 的损失函数为 soft dice loss 与 cross-entropy loss，如下等式所示：

$$L(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j} log Y_{i,j} \tag{1}$$

其中 I 为体素的数量，J 是类别数，$Y_{i,j}$ 和 $G_{i,j}$ 表示类别 j 在体素 i 输出概率和真实类别独热编码。

IV. Experiment

这一部分将会介绍实验数据，实验设置以及展示实验结果。

A. Dataset

本次实验的 3D 图像数据来源于暨南大学附属第一医院（非公开数据）。总共分为两个类型的 CT 图像数据：平扫 CT 图像与增强 CT 图像。这里简单描述一下平扫 CT 图像与增强 CT 图像的区别：增强 CT 图像能更好的显示肿瘤区域，但并不是所有病人都适合使用增强 CT 图像进行诊断。采集样本数量总共为 184，且每一个样本均有对应的平扫 CT 图像与增强 CT 图像，所以全部样本数量应为 184×2，每个样本的两种图像序列均有对应的肿瘤勾画掩膜。数据均为 dcm 格式文件，每个文件中包含多张切片，称之为序列，每一张切片的分辨率为 512×512。样本示例如图 4所示。理论上可单独选取平扫或增强 CT 图像进行训练，或者二者一起使用效果更佳。但由于处理 3D 图像所需内存较大，且训练时间较长，这里仅选取平扫 CT 图像进行实验，随机选取 80% 作为训练集，其余 20% 作为测试集。

B. Image Transforms

在输入模型之前，我们需要对输入图像进行一定的变换，作用相当于数据增强。由于图像序列来源于不同型号的扫描机器，得到的图像数据具体参数不完全一致，所以首先对图像进行重采样到统一的分辨率。然后选取我们所需的信号值——食管组织对应的 HU 值，将其缩放至 $[0, 1]$ 范围。由于单张切片较大，我们可以选择对切片进行分割得到子图，这里选择分割为 4 张子图。另外，还可对图像进行随机旋转，缩放等操作来进行图像增强，提高模型泛化性能。

C. Experimental Settings

实验选择了两个模型：传统的 3D Unet 与较新的基于 Transformer 的 UNETR 来进行对比。两个模型的的基本参数设置如图 5所示。损失函数选择的是 FocalLoss，FocalLoss 的定义如式 2所示。选择 Adam 作为优化器。另外，所选取的评价指标为：Dice score(DSC)，Hausdorff distance(HD) 和 Average Surface Distance(ASD)。最大迭代次数设置为 500。

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t),$$
$$p_t = \begin{cases} p & \text{if y=1} \\ 1-p & \text{otherwise} \end{cases} \tag{2}$$

D. Experimental Results

表 I中展示了本次实验的结果。

```
model = UNet(
        dimensions=3,
        in_channels=1,
        out_channels=2,
        channels=(16, 32, 64, 128, 256),
        strides=(2, 2, 2, 2),
        num_res_units=2,
        norm=Norm.BATCH,
).to(device)
```

(a) 3D Unet 模型基本参数

```
model=UNETR(
        in_channels=1,
        out_channels=2,
        img_size=(96,96,64),
        feature_size=16,
        hidden_size=768,
        mlp_dim=3072,
        num_heads=12,
        pos_embed="conv",
        norm_name="instance",
        res_block=True,
        dropout_rate=0.0
).to(device)
```

(b) UNETR 基本参数设置

图 5. 3D Unet 与 UNETR 基本参数设置

1) 3D U-Net: 我们首先采用了 FocalLoss 损失函数进行训练，记为模型 3D U-Net-1。实验结果如下：训练损失下降与 DSC 评价指标变化如图 6所示。由于距离计算问题，训练过程中 HD 与 ASD 一直为 Inf 值，无实质意义，故不予展示。500 个 epoch 中最优的 DSC 出现在第 450 个 epoch，值为 0.5247。可见距离 DSC 目标值 0.8 还有较大距离。然后我们采用 DiceFocalLoss 损失函数进行进一步尝试，记为模型 3D U-Net-2。训练损失下降与评价指标变化如图 7所示。500 个 epoch 中最优的 DSC 出现在第 210 个 epoch，值为 0.5353。此时的 HD 为 48.7670mm，ASD 为 7.4778mm。

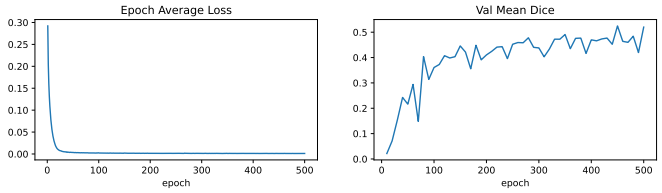


图 6. 3D U-Net-1 训练集损失与测试集 DSC 变化图

2) UNETR: 对于 UNETR 模型，我们首先采用了 DiceLoss 损失函数进行训练，记为 UNETR-1。UNETR-1 的训练损失下降与评价指标变化如图 8所示。500 个 epoch 中最优的 DSC 出现在第 330 个 epoch，值为 0.4684。此时的 HD 为 213.1135mm,ASD 为 29.95mm。结果仍是不理想。接着采用 DiceFocalLoss 损失函数进行尝试，这里 DiceFocalLoss 是 DiceLoss 与 FocalLoss 的加权和，记作 UNETR-2。UNETR-2 的训练损失下降与评价指标变化如图 9所示。500 个 epoch 中最优的 DSC 出现在第 450 个 epoch，值为 0.4694。此时的 HD 为 122.3268mm，ASD 为 12.5142mm。虽然 DSC 值提升不大，但是其他两个评价指标都大幅下降，说明了更换损失函数具有一定的效果提升。另外

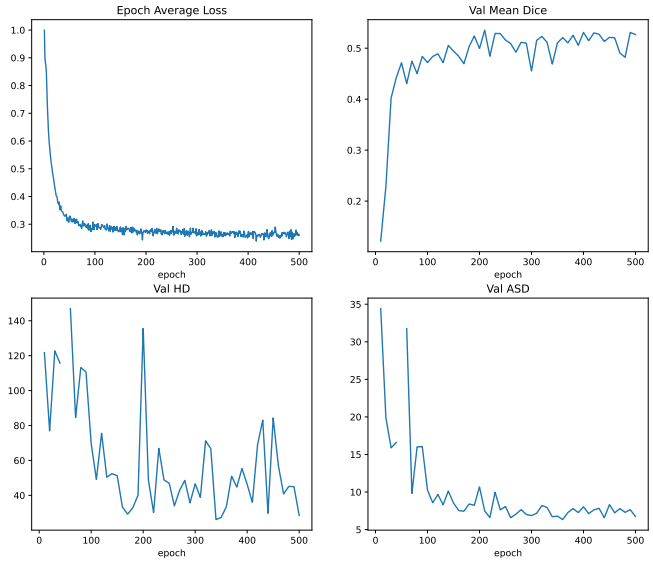


图 7. 3D U-Net-2 训练集损失与测试集评价指标变化图

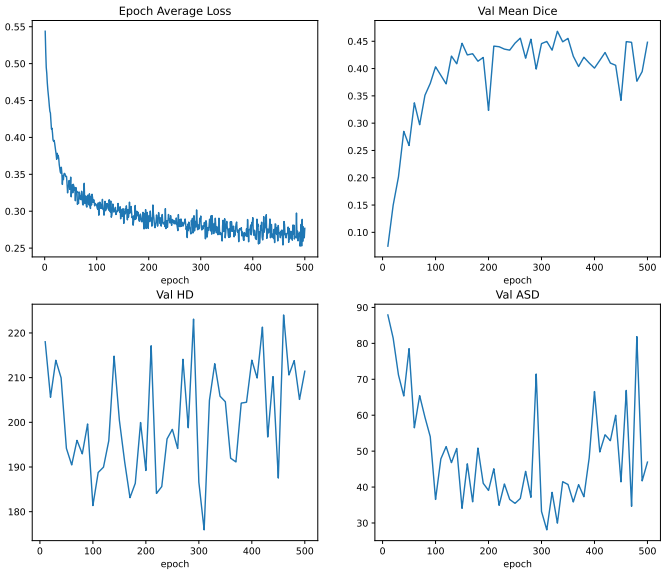


图 8. UNETR-1 训练集损失与测试集评价指标变化图

V. Conclusion

本次实验采用了 3D U-Net 与 UNETR 完成食管癌肿瘤 3D 分割的任务，但遗憾的是未能达到预期的结果。从上述结果中，我们可以发现：SOTA 模型所取得的 DSC 值并不高，我们所采用两个模型的 DSC 值也都不高，表明食管癌 GTV 分割任务难度确实比较大，这反映了本次任务本身非常具有挑战性。

从我们所采用的两个模型的表现来看：传统的 3D U-Net 表现得比较好，损失下降更快且收敛稳定，模型

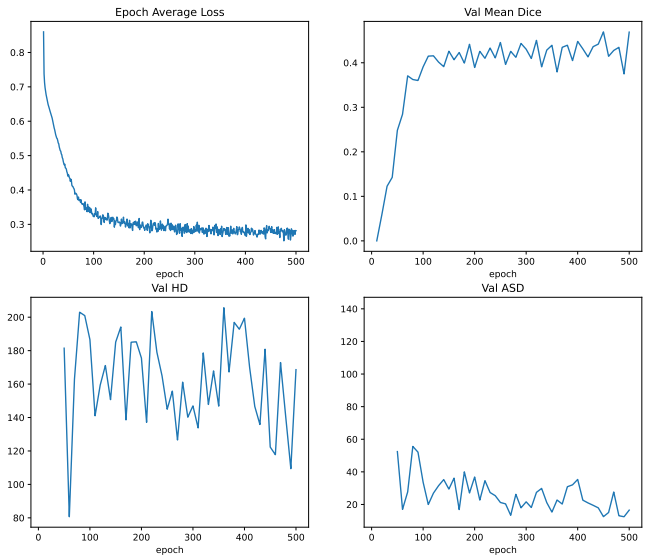


图 9. UNETR-2 训练集损失与测试集评价指标变化图

表 I
平均 DSC，HD 和 ASD，前两个模型都是 SOTA。3D U-Net-1 使用的是 FocalLoss 损失函数，3D U-Net-2 使用的是 DiceFocalLoss 损失函数，UNETR-1 使用的是 DiceLoss 损失函数，UNETR-2 使用的是 DiceFocalLoss 损失函数

methods	DSC	HD(mm)	ASD(mm)
DenseUNet [2]	0.703	72.2	14.2
PSNN [2]	0.751	43.7	6.7
3D U-Net-1	0.5247	inf	inf
3D U-Net-2	0.5353	48.7670	7.4778
UNETR-1	0.4684	213.1135	29.95
UNETR-2	0.4694	122.3268	12.5142

较为简单有效；而基于 Transformer 实现的 UNETR 的表现并未达到预期，这有可能是因为基于 Transformer 实现的模型一般都需要在比较大的数据上进行预训练，然后进行迁移学习来达到比较好的学习效果。

本次实验是一个尝试性的实验，尽管结果并未如人意。但是也给我提供了一些改进的思路：首先是数据量的问题。本次实验由于时间和计算资源的限制，并未将全部数据投入使用，仅仅使用了其中一种 CT 图像序列。前面提到，使用增强的 CT 图像序列有可能效果更佳，或者将数据全部投入使用，采用交叉验证的方式。这些都可以作为下一步的改进方案。然后是模型的问题。本次实验仅采用了两个模型来进行尝试，这点是不太足够的。下一步可尝试更多的模型，或者尝试自己基于现有模型提出改进的模型。再其次就是一些调参优化的问题。本次实验也尝试了更换损失函数的方式来进行调参，结果也体现了这一举措的有效性。进一步我们可以增加图像增强的方式，动态调整学习率等等，当然，这些 tricks 都必须建立在模型有效能实现较好表现的基础之上。

虽然本次课程作业到这里就结束了，但是这个 3D 分割任务还仅仅是个开始，之后仍可以基于这个任务开展许多工作与探究。

参考文献

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-Kan Tseng, and Le Lu. Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Medical Image Analysis*, 68:101909, 2021.

[3] Sahar Yousefi, Hessam Sokooti, Mohamed S Elmahdy, Irene M Lips, Mohammad T Manzuri Shalmani, Roel T Zinkstok, Frank JWM Dankers, and Marius Staring. Esophageal tumor segmentation in ct images using a dilated dense attention unet (ddaunet). *IEEE Access*, 9:99235–99248, 2021.

[4] Shan Tan, Laquan Li, Wookjin Choi, Min Kyu Kang, Warren D D' Souza, and Wei Lu. Adaptive region-growing with maximum curvature strategy for tumor segmentation in 18f-fdg pet. *Physics in Medicine & Biology*, 62(13):5383, 2017.

[5] Tobias Fechter, Sonja Adebahr, Dimos Baltas, Ismail Ben Ayed, Christian Desrosiers, and Jose Dolz. A 3d fully convolutional neural network and a random walker to segment the esophagus in ct. *arXiv preprint arXiv:1704.06544*, 2017.

[6] Roger Trullo, Caroline Petitjean, Dong Nie, Dinggang Shen, and Su Ruan. Fully automated esophagus segmentation with a hierarchical deep learning approach. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 503–506. IEEE, 2017.

[7] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-kan Tseng, and Le Lu. Deep esophageal clinical target volume delineation using encoded 3d spatial context of tumors, lymph nodes, and organs at risk. In *International conference on medical image computing and computer-assisted intervention*, pages 603–612. Springer, 2019.

[8] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-Kan Tseng, and Le Lu. Accurate esophageal gross tumor volume segmentation in pet/ct using two-stream chained 3d deep network fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 182–191. Springer, 2019.

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021.

附录

另外我还利用增强 CT 序列进行了初步的尝试，但由于实验尚未成体系，在此就以附录的形式做一个记录。

我们依然采用 3D U-Net 模型，选择 DiceFocalLoss 作为损失函数，但只训练 300 个 epoch。训练损失下降与评价指标变化如图 10所示。

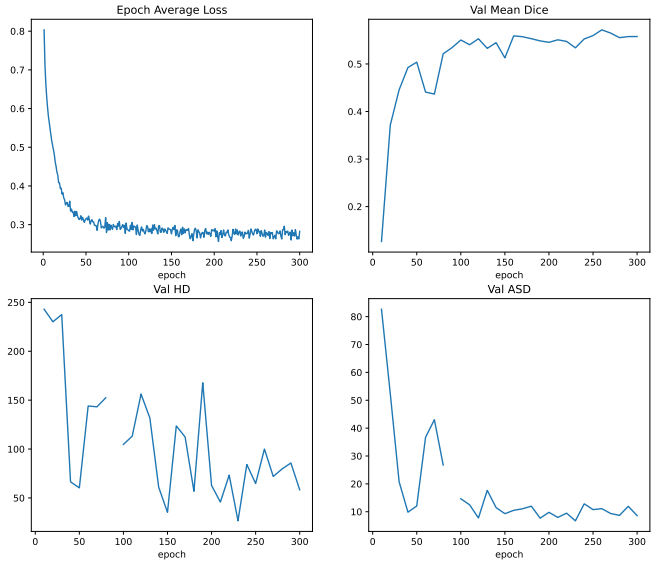


图 10. 3D U-Net 在增强 CT 图像上训练集损失与测试集评价指标变化图

300 个 epoch 中最优的 DSC 出现在第 260 个 epoch，值为 0.5717。此时的 HD 为 99.9832mm，ASD 为 11.1091mm。可见使用增强 CT 图像来完成食管癌 GTV 分割任务确实会有所提升。后续可利用增强 CT 图像逐步补充实验来实现较好的 GTV 分割效果。