实验细节：

一．复现结果与对比

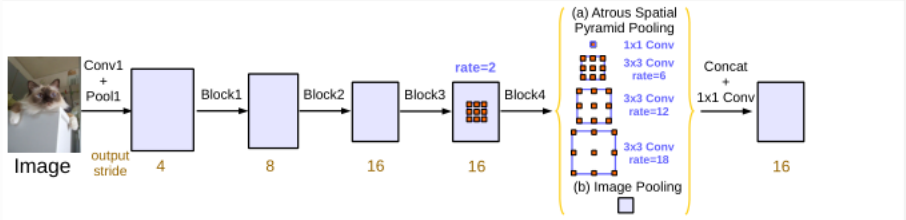
本篇文章的工作主要分为两个部分：对左心室进行帧级语义分割、根据原始超声心动图预测射血分数，以下分别展开叙述。

（一） 左心室语义分割

语义分割任务采取弱监督学习方式进行训练。训练数据中，每个超声心电图视频带有左心室分割标签，这些标签由人类专家追踪左心室，在每个视频特定的两个时间点（收缩末期和舒张末期）手动分割标记左心室而产生。模型经过学习后，可以对原始超声心动图的每一帧进行左心室语义分割。

1. 网络结构

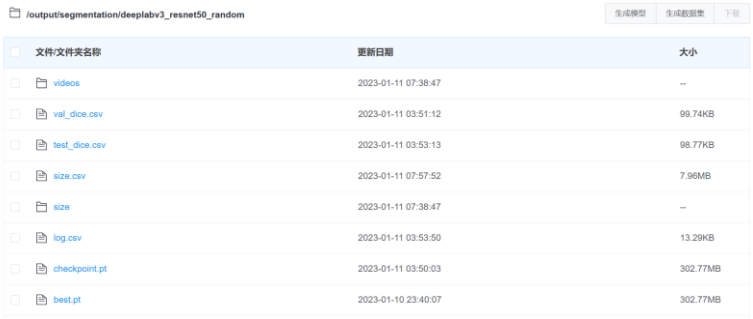
左心室分割任务采用经典的语义分割模型 DeepLabv3，主干的特征提取部分(Block1~4)使用 resnet50 网络。DeepLabv3 网络的主要特点是引入空洞卷积（atrous convolutions）来解决随着下采样进行使特征分辨率逐渐降低的问题，并且引入 ASPP，ASPP 中不同的空洞率可以有效捕获不同尺度的信息。如图所示：



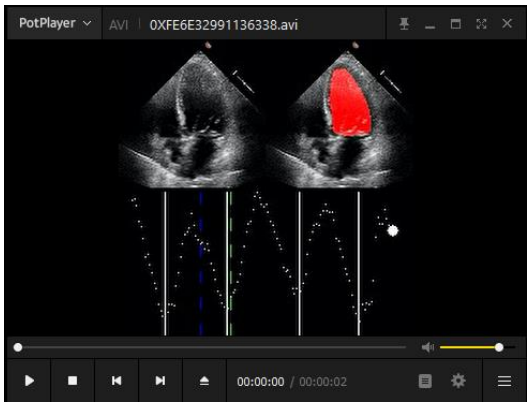
2. 实验设计

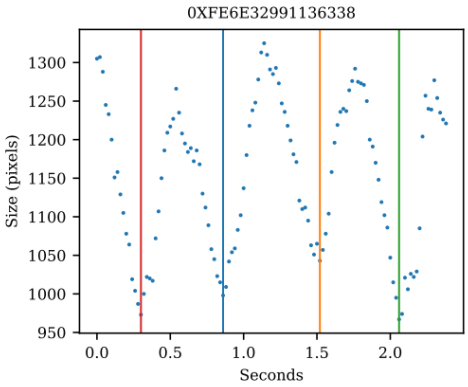
除 deeplabv3_resnet50 模型以外，另外设置三个对比试验：1. 将 backbone 网络替换为 resnet101，即 deeplabv3_resnet101，2.3：使用另一种经典语义分割网络 FCN 替代 DeepLabv3,分别组成 fcn_resnet50,fcn_resnet101. 共四组实验

每组实验的输出如图所示：



/output/videos 保存了测试集中所有超声心动图的分割结果：（左图）





/output/size 保存了测试集中所有超声心动图的左心室体积变化图：（右图）

Log.csv 记录了训练过程中训练集和验证集的 loss、Dice 系数变化情况，test_dice.csv 和 val_dice.csv 分别记录了测试集和训练集中每个视频的 Dice 系数。

a. Deeplabv3_resnet50

Best validation loss 0.03569907191014589 from epoch 15
val dice (overall): 0.9222 (0.9207 - 0.9237)
val dice (large): 0.9074 (0.9051 - 0.9096)
val dice (small): 0.9316 (0.9296 - 0.9335)
test dice (overall): 0.9230 (0.9216 - 0.9243)
test dice (large): 0.9080 (0.9056 - 0.9102)
test dice (small): 0.9325 (0.9310 - 0.9340)

b. Deeplabv3_resnet101

Best validation loss 0.040438333562850225 from epoch 34
val dice (overall): 0.9130 (0.9112 - 0.9148)
val dice (large): 0.8960 (0.8935 - 0.8986)
val dice (small): 0.9238 (0.9214 - 0.9261)
test dice (overall): 0.9146 (0.9133 - 0.9161)
test dice (large): 0.8974 (0.8949 - 0.8997)
test dice (small): 0.9256 (0.9241 - 0.9272)

c. Fcn_resnet50

Best validation loss 0.039766328025640896 from epoch 48
val dice (overall): 0.9143 (0.9125 - 0.9160)
val dice (large): 0.8989 (0.8965 - 0.9013)
val dice (small): 0.9241 (0.9217 - 0.9264)
test dice (overall): 0.9162 (0.9148 - 0.9176)
test dice (large): 0.9005 (0.8981 - 0.9029)
test dice (small): 0.9262 (0.9245 - 0.9279)

d. Fcn_resnet101

Best validation loss 0.03697809708830158 from epoch 34
val dice (overall): 0.9178 (0.9162 - 0.9193)
val dice (large): 0.9032 (0.9010 - 0.9054)
val dice (small): 0.9271 (0.9249 - 0.9290)
test dice (overall): 0.9187 (0.9174 - 0.9201)
test dice (large): 0.9040 (0.9018 - 0.9061)
test dice (small): 0.9282 (0.9266 - 0.9297)

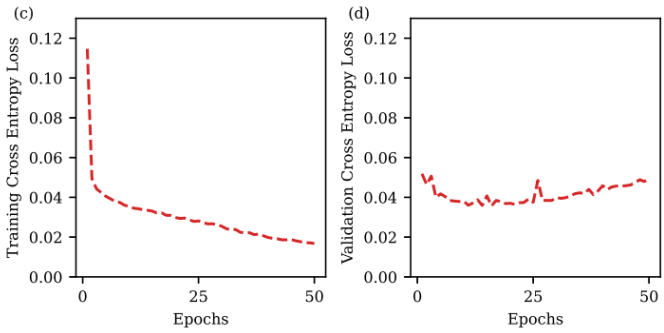
对比分析：

四组对比试验采用相同参数设置：num_epochs=50.根据日志文件记载，val_loss 均在 50 个 epoch 之前达到最小值，然后出现过拟合，说明训练充分。对比四组实验的 val_loss 及 Dice 系数，a 组实验：DeepLabv3_resnet50，即文中采用的体系结构性能最优。

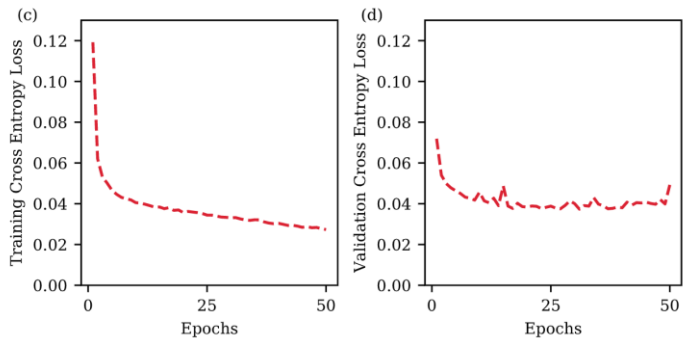
可能原因分析：DeepLabv3 结构比 fcn 具有优越性，前文已提到 DeepLabv3 的主要贡献，在此不再赘述；在语义分割任务中，上下文之间的相互作用以及空间信息特征有利于准确定位。Resnet101 相比 resnet50 层数更深，提取到的是更高层次的语义特征，而空间信息保存相对更少。且输入尺寸为 112x112 较小，在经过 resnet 的每一层时需要进行 padding 操作来保持输入输出特征维度一致，resnet 层数过多可能会对数据分布产生影响。

3. 性能对比

画出最优 DeepLabv3_resnet50 组训练过程中的 loss 曲线：



文中对应训练曲线：



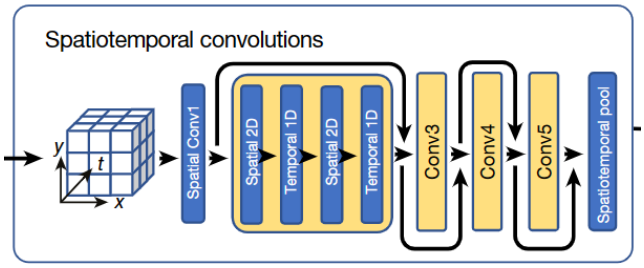
DeepLabv3_resnet50 组在测试集上的 Dice 系数达到 0.9230，与文中所说的 0.92 保持一致。综上，左心室语义分割部分取得了较好的复现效果。

（二） 射血分数预测

将射血分数预测任务作为分类任务进行学习。

1. 网络结构

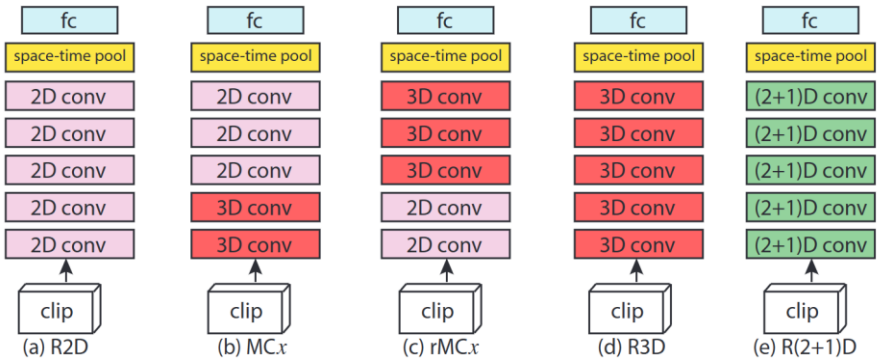
其网络的主要特征是使用了 3D 卷积和带有残差连接的 CNN 结构。与以往用于医学图像处理的二维卷积不同，时空卷积结合了二维的空间信息以及第三维的时间信息，被广泛应用在视频动作识别与分类任务中。文章测试并对比了三种具有可变时间卷积积分的模型体系结构（R3D、MC3 和 R2+1D），最终选择 R2+1D 作为 Echonet-Dynamic 使用的具有最佳性能的体系结构。



2. 实验设计

在本篇文章的对比实验中，主要考虑三个方面的因素：

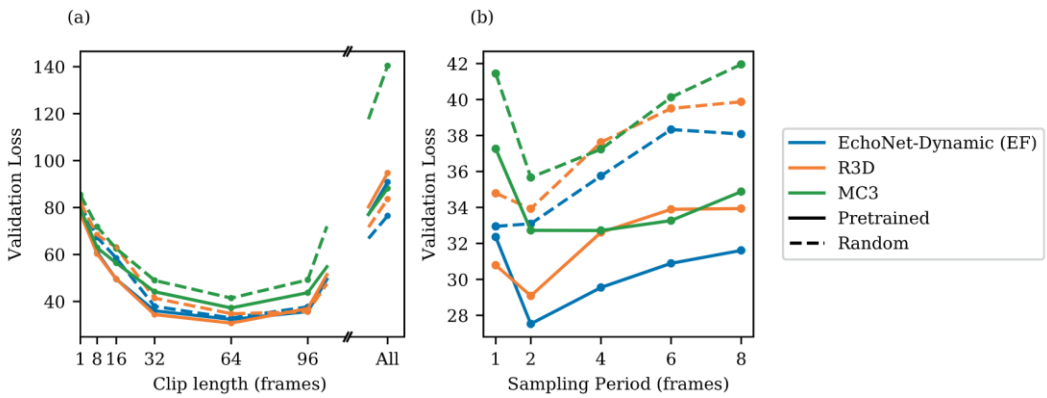
a. 测试三种具有可变时空卷积积分的模型体系结构（R3D、MC3 和 R2+1D）.在 R3D 结构中，所有卷积层联合考虑空间和时间维度，这些卷积层由五个卷积块组成。MC3 和 R2+1D 架构是作为只考虑空间关系的二维卷积和 R3D29 使用的全三维卷积之间的中间基础而引入的。MC3 架构将最后三个块中的卷积替换为二维卷积，R2+1 架构显式地将所有三维卷积分解为二维空间卷积，然后是一维时间卷积。结构图如图所示：



b. 时空卷积模型使用 Kinetics-400 数据集的预训练模型的权重进行初始化或使用随机参数初始化

c. clip_length 和 period 决定了如何对原始超声心电图进行采样和剪裁，作为模型的输入。对这两个参数进行了 hyperparameter_sweep，确定不同参数对模型性能的影响。（clip_length 可取值 1,8,16,32,64,96 和 all frames ; period 可取值 1,2,4,6,8）

综合考虑以上因素，一共需要设置 70 组对比试验。因为实验组别过多，考虑复现时间以及算力的限制，对于 c，我直接使用了文章实验得出的结论：frame=32,period=2 时模型性能最佳，如下图所示：



并在此基础上，对 a、b 两组条件设置对照试验，共 6 组:

a. r2+1d_pretrained

Best validation loss 27.761623465496562 from epoch 32
val (one clip) R2: 0.808 (0.786 - 0.828)
val (one clip) MAE: 4.07 (3.91 - 4.23)
val (one clip) RMSE: 5.39 (5.16 - 5.62)
val (all clips) R2: 0.827 (0.807 - 0.845)
val (all clips) MAE: 3.90 (3.74 - 4.05)
val (all clips) RMSE: 5.11 (4.89 - 5.33)
test (one clip) R2: 0.788 (0.763 - 0.810)
test (one clip) MAE: 4.20 (4.03 - 4.37)
test (one clip) RMSE: 5.63 (5.37 - 5.89)
test (all clips) R2: 0.809 (0.787 - 0.828)
test (all clips) MAE: 4.00 (3.84 - 4.17)
test (all clips) RMSE: 5.35 (5.10 - 5.61)

b. r2+1d_random

Best validation loss 32.619324156956644 from epoch 41
val (one clip) R2: 0.773 (0.746 - 0.797)
val (one clip) MAE: 4.34 (4.16 - 4.52)
val (one clip) RMSE: 5.86 (5.59 - 6.13)
val (all clips) R2: 0.798 (0.775 - 0.819)
val (all clips) MAE: 4.20 (4.04 - 4.37)
val (all clips) RMSE: 5.52 (5.28 - 5.76)
test (one clip) R2: 0.755 (0.726 - 0.781)
test (one clip) MAE: 4.43 (4.24 - 4.62)
test (one clip) RMSE: 6.05 (5.75 - 6.35)
test (all clips) R2: 0.777 (0.749 - 0.801)
test (all clips) MAE: 4.24 (4.06 - 4.42)
test (all clips) RMSE: 5.78 (5.49 - 6.07)

c. mc3_pretrained

Best validation loss 32.80200073289575 from epoch 40
val (one clip) R2: 0.773 (0.748 - 0.795)
val (one clip) MAE: 4.45 (4.28 - 4.62)
val (one clip) RMSE: 5.87 (5.60 - 6.13)
val (all clips) R2: 0.794 (0.772 - 0.813)
val (all clips) MAE: 4.24 (4.07 - 4.41)
val (all clips) RMSE: 5.59 (5.34 - 5.83)
test (one clip) R2: 0.769 (0.743 - 0.792)
test (one clip) MAE: 4.44 (4.27 - 4.62)
test (one clip) RMSE: 5.87 (5.61 - 6.15)
test (all clips) R2: 0.789 (0.766 - 0.810)
test (all clips) MAE: 4.23 (4.06 - 4.40)
test (all clips) RMSE: 5.61 (5.35 - 5.89)

d. mc3_random

Best validation loss 36.06559305605681 from epoch 43
val (one clip) R2: 0.747 (0.721 - 0.771)
val (one clip) MAE: 4.65 (4.46 - 4.84)
val (one clip) RMSE: 6.19 (5.92 - 6.45)
val (all clips) R2: 0.778 (0.754 - 0.798)
val (all clips) MAE: 4.41 (4.24 - 4.59)
val (all clips) RMSE: 5.80 (5.57 - 6.04)
test (one clip) R2: 0.726 (0.696 - 0.753)
test (one clip) MAE: 4.73 (4.54 - 4.94)
test (one clip) RMSE: 6.40 (6.10 - 6.70)
test (all clips) R2: 0.752 (0.725 - 0.777)
test (all clips) MAE: 4.53 (4.35 - 4.73)
test (all clips) RMSE: 6.09 (5.80 - 6.39)

e. r3d_pretrained

Best validation loss 28.727937532507855 from epoch 16
val (one clip) R2: 0.811 (0.788 - 0.831)
val (one clip) MAE: 4.05 (3.90 - 4.22)
val (one clip) RMSE: 5.35 (5.12 - 5.59)
val (all clips) R2: 0.826 (0.805 - 0.843)
val (all clips) MAE: 3.91 (3.76 - 4.07)
val (all clips) RMSE: 5.14 (4.92 - 5.35)
test (one clip) R2: 0.777 (0.752 - 0.799)
test (one clip) MAE: 4.30 (4.12 - 4.48)
test (one clip) RMSE: 5.77 (5.51 - 6.05)
test (all clips) R2: 0.803 (0.780 - 0.823)
test (all clips) MAE: 4.05 (3.88 - 4.22)
test (all clips) RMSE: 5.43 (5.17 - 5.70)

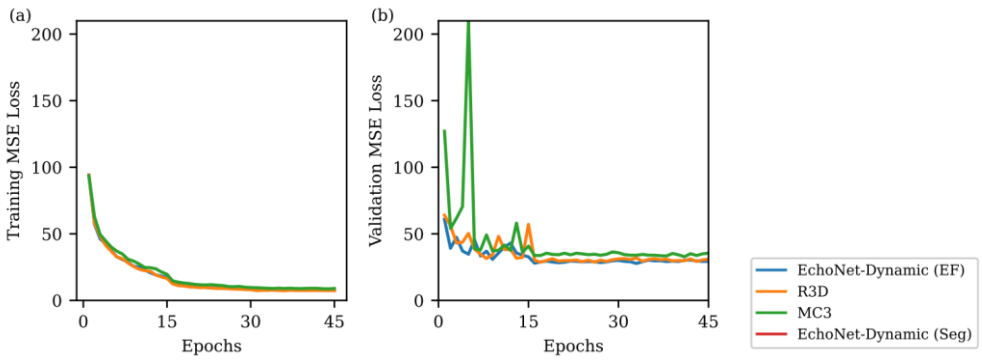
f. r3d_random

Best validation loss 32.33228710719517 from epoch 33
val (one clip) R2: 0.781 (0.755 - 0.803)
val (one clip) MAE: 4.34 (4.17 - 4.52)
val (one clip) RMSE: 5.76 (5.51 - 6.01)
val (all clips) R2: 0.797 (0.774 - 0.817)
val (all clips) MAE: 4.21 (4.04 - 4.37)
val (all clips) RMSE: 5.54 (5.31 - 5.77)
test (one clip) R2: 0.750 (0.721 - 0.776)
test (one clip) MAE: 4.49 (4.30 - 4.68)
test (one clip) RMSE: 6.12 (5.82 - 6.41)
test (all clips) R2: 0.779 (0.752 - 0.802)
test (all clips) MAE: 4.27 (4.10 - 4.45)
test (all clips) RMSE: 5.75 (5.47 - 6.05)

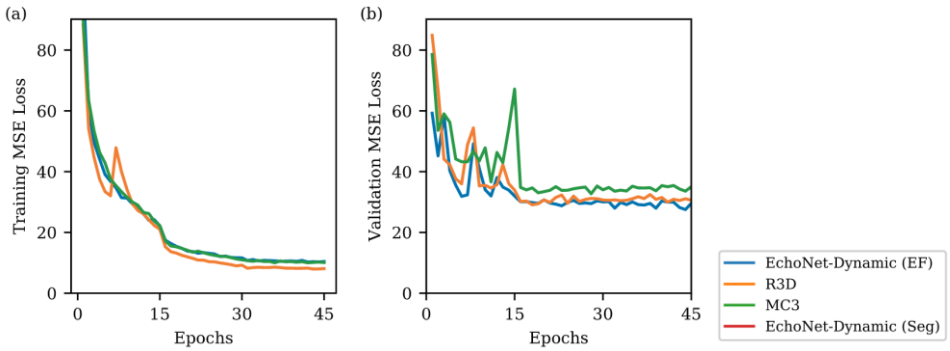
从以上对照试验可以看出，使用预训练模型进行参数初始化可以提高模型的性能。并且在三种时空卷积模型结构中，r2+1d 的性能最好，这与文章得出的结论一致。

3. 性能对比

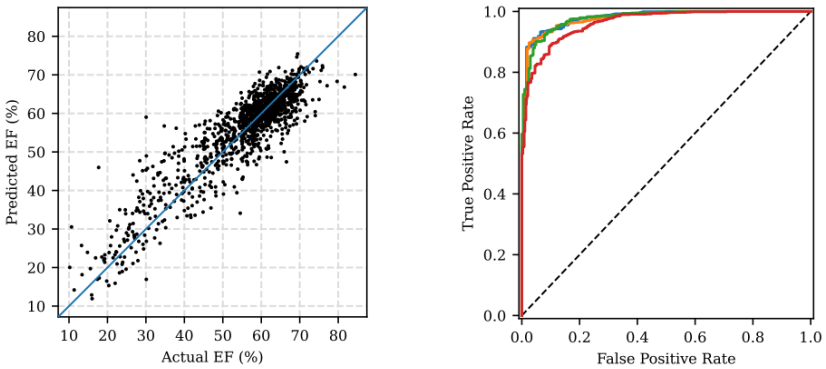
绘制性能最优的模型：r2+1d_pretrained 的训练 loss 曲线：



与文章展示的训练曲线对比：



另外，分别绘制测试集上射血分数标签与预测值的散点图与模型的 ROC 曲线：

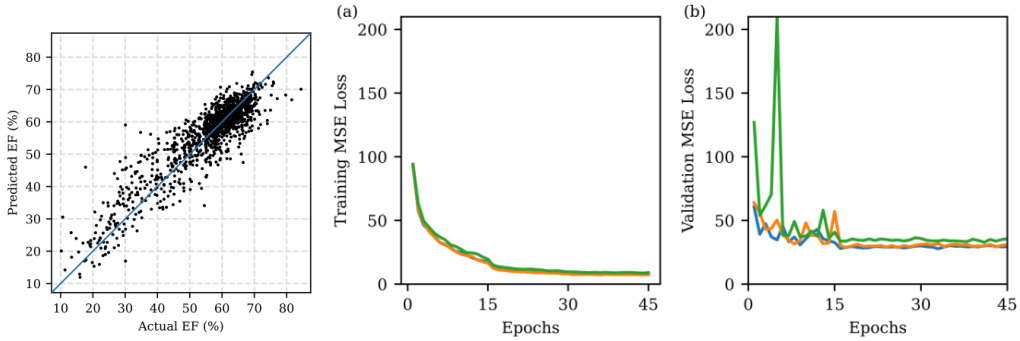


其中标签与预测值的决定性系数R^2达到 0.788（one clip）和 0.809（all clips），这与文章中的实验结果 0.81 较接近。选择四种不同阈值（35、40、45、50）画出的 ROC 曲线，AUC 值如下图所示，与文章中提到的 0.97 较接近。即模型能较好的将心肌衰弱（射血分数偏低）的样本分类出来。综上所述，射血分数预测模型得到了较好的复现。

```
35 0.9796496975806451
40 0.9779681899641578
45 0.9781469490781387
50 0.963534972648574
```

二．网络改进

由于文章使用的网络都是 torchvision.models 中封装的经典网络模型，没有对网络本身进行改动。



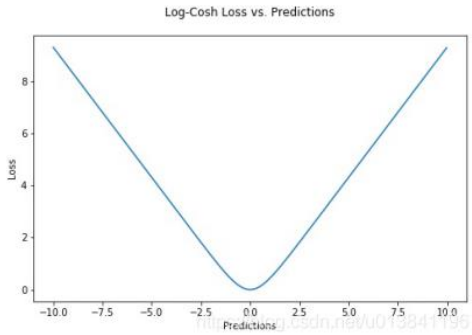
观察到预测射血分数任务中，从散点图可以看出，当真实 EF 的值较小时，有一些点的预测值与 groud_truth 的偏差较大。并且观察验证集上的 loss 变化曲线，发现有一些比较大的抖动。尝试使用其他形式的 loss 函数来更换简单的 MSE 损失，来观察对实验结

果的影响。

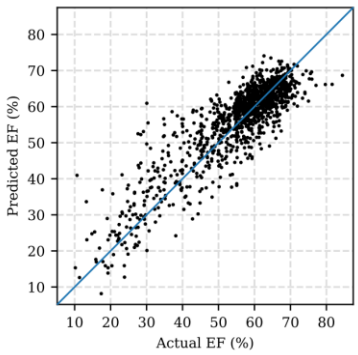
1. Log_cosh 函数:

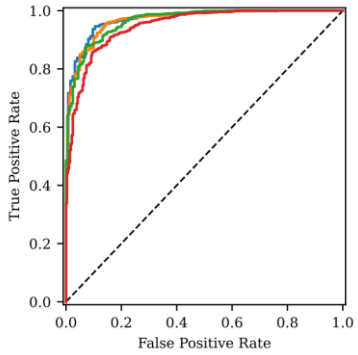
$$L(y, y^p) = \sum_{i=1}^{n} \log(\cosh(y_i^p - y_i))$$

对于较小的 X 值，log(cosh(x))约等于(x ** 2) / 2；对于较大的 X 值，则约等于 abs(x) - log(2)。这意味着 Log_cosh 很大程度上工作原理和平均方误差很像，但偶尔出现错的离谱的预测时对它影响又不是很大。它具备了 Huber 损失函数的所有优点，但不像 Huber 损失，它在所有地方都二次可微。在 X 较小时，接近 MSE，在 X 较大时，接近 MAE。



结果:





Best validation loss 3.9609330633412236 from epoch 19

val (one clip) R2:	0.739 (0.708 - 0.767)
val (one clip) MAE:	4.58 (4.38 - 4.78)
val (one clip) RMSE:	6.28 (5.97 - 6.59)
val (all clips) R2:	0.760 (0.732 - 0.786)
val (all clips) MAE:	4.46 (4.27 - 4.64)
val (all clips) RMSE:	6.03 (5.73 - 6.32)
test (one clip) R2:	0.731 (0.700 - 0.758)
test (one clip) MAE:	4.73 (4.54 - 4.92)
test (one clip) RMSE:	6.35 (6.05 - 6.65)
test (all clips) R2:	0.760 (0.734 - 0.784)
test (all clips) MAE:	4.44 (4.26 - 4.63)
test (all clips) RMSE:	5.99 (5.70 - 6.28)

与文章结果相比，没有明显的变化，R^2分别为 0.731 (one clip) 和 0.760 (all clips)，略有下降。

除了 log_cosh 损失之外，我发现回归问题的几种常见损失函数中，例如 MSE、MAE，都属于绝对误差，但实际上预测值和真实值之间相同的偏差值，对于不同的真实值，他们所代表的偏差程度并不相同，于是尝试两种相对误差损失，分别为 MAPE 和 sMAPE，定义式如下。但经过实际测试，得到了很差的实验结果，暂时还没有想到比较合理的解释。由于篇幅限制，在此不详细展开讨论。

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad \mathrm{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2}$$

三．实验总结

1. 实验中遇到的困难
   a. 在本篇文章中，集成了大量的新概念。包括弱监督学习、语义分割中标签的复杂性（使用线段来覆盖标记目标区域）、Dice 系数、ROC 曲线、三维时空卷积等深度学习知识，以及心脏功能和临床治疗等相关的生物知识，这对复现过程中对代码和文章的理解造成了一定的困难。
   b. 文章中提到，所有模型的验证测试均在一张 GTX1080Ti 显卡上完成。在实际复现中，与文章使用相同训练参数的情况下（batch_size,frame,period），同时使用两张 GTX1080Ti 才能满足需求。另外，使用官方提供的训练代码，经常会因内存溢出（注意不是显存）而造成训练进程被意外终结。经过反复对比查找，发现原始代码中的两个参数设置会使内存溢出。
   num_worker：原始代码将其值设置为 4，目的是使用多个子进程进行数据加载进而加快训练速度，这与 CPU 有关而与 GPU 无关，所以设置过大容易造成内存溢出。将其修改为 1.
   pin_memory:是 torch.utils.data.DataLoader()一个参数。pin_memory 就是锁页内存，创建 Dataloader 时，设置 pin_memory=True，则意味着生成的 Tensor 数据最开始是属于内存中的锁页内存，这样将内存的 Tensor 转移到 GPU 的显存就会更快一些。主机中的内存，有两种存在方式，一是锁页，二是不锁页。锁页内存存放的内容在任何情况下都不会与主机的虚拟内存进行交换，而不锁页内存在主机内存不足时，数据会存放在虚拟内存中。当计算机的内存充足时，可以设置 pin_memory=True，但当系统交换内存使用较多时，应设置 pin_memory=False,并与主机的硬件性能相关。源代码中将 pin_memory 设置为 device.type==”cuda”,即将其设置为 True，容易造成内存溢出。将其修改为默认值：False.
2. 实验心得

   通过本次作业，我对语义分割这一领域有了较多新的认识和了解，比如经典网络（FCN、空洞卷积、DeepLab 系列以及 ASPP 等）和评价指标（Dice 系数）。

   对深度学习的“科学性”加深了认识：之前总觉得深度学习是一个非常玄学和只依赖经验的概念，在经过设置多种对照试验过程中，逐渐体会到其严谨性和科研的大致流程。

   本篇文章属于 Nature 期刊而非会议论文，且与生物有较大的交叉性，给我的最大感受是，真正的将深度学习作为一种手段，去解决实际生活中的问题，而不是单纯的刷 banchmark(当然也有其价值)。让我看到了从事人工智能领域的最终意义所在。