第 3 章　增强超声视频数据迁移学习探讨

3.1　引言

本章探讨自然数据集的模型在 CEUS 中的迁移性能，数据来自中山大学第一附属医院超声科，由医生回顾性收集 1999-2019 年中山大学附属第一医院胆管细胞癌（intrahepatic cholangiocellular carcinoma, ICC）患者。入组标准：1、病理确诊；2、病人进行超声造影检查；3、无肝外转移、血管侵犯；4、超声造影成像前未进行过系统或局部治疗。共 190 例胆管细胞癌患者入组，为胆管细胞癌患者按 1：1 比例匹配肝细胞癌 (（hepatocellular carcinoma，HCC）病例，所有胆管细胞癌和肝细胞癌病例组成本课题的研究病例，共计 380 人，研究采用三折的交叉验证，按肿瘤类型随机分组。入组数据为 CEUS 前 120s 的视频，使用 RandiAnt 转出为 “.bmp” 格式的图像序列，输入网络时只将需要的图片读入再合并为一个 clip，从而降低数据读取时的内存和加快读取速度，所有病例由影像医生挑选和人工勾画了以下 8 张图片的肿瘤靶区：1、病变最大切面动脉增强早期图像；2、病变最大切面动脉增强中期图像；3、病变最大切面动脉增强峰值图像；4、增强峰值后 5s 图像；5、增强峰值后 10s 图像；6、增强峰值后 20s 图像；7、增强峰值后 40s 图像；8、CEUS120s 的图像。通过对一个视频中的 8 个 Bounding-box 取最小的能够包下所有区域的 Bounding-box，并将其上下左右向外扩展 20 个像素获取单个视频统一的截取区域，从而实现病灶区域和背景分离。

肝细胞癌占据成人原发性肝癌的 90%(Galle 等, 2018)；胆管细胞癌是肝内第二大高发癌症，恶性程度明显高于 HCC，数据也更为稀有，病人的长期存活只能通过手术切除来实现，在临床上必须明确与肝细胞癌区分。2017 年，大规模多中心的研究Aube 等 (2017) 表明，超声造影在 HCC 的诊断敏感度较低，对于病灶大小在 10-20mm 的病灶仅为 39.6%（CT 63.5%，MRI 70.6%），对于 20-30mm 的病灶敏感度为 52.9%（CT 71.6%, MRI 72.3%），存在和胆管细胞癌混淆的风险。2018 年，多中心研究Terzi 等 (2018) 对于 1006 个结节细化分析后表明，对于 LI-RADS LR5 期的病人，超声造影的阳性预测值可以达到 99%，虽然在这 1006 个病人中只有 519 个病人可以归类为 LR5 期病人。Leoni 等 (2013) 指出，相比 CT 和 MRI，CEUS 的消退表征对于小于 20mm 的病灶不易检测，因而敏感度较低。本次研究

入住的病例没有将病灶尺寸作为入组的筛选条件，是希望模型的通用性更好，本研究为二分类问题，HCC 标签值为 0，ICC 标签值为 1。

由于大规模数据集上训练视频网络是不能靠一台服务器和几张显卡实现的任务，本次研究使用公开发表的模型和参数，检索依据来自 Paper With Code 网站上计算机视觉-> 视频-> 视频分类任务的排名表，部分公开模型虽然有着较好的性能，但因为作者认为这些网络的计算成本太高，如 facebook 开源的视频网络模型库VMZ中的 ir-CSN-152 和 ir-CSN-152 模型，输入 32×224^2, 152 层深度，没有纳入实验中。

本研究对比了 2D CNN，2D CNN+BERT，C3D，I3D，SlowFast，R(2+1)D，TSM，GSM 后，得到结论如下：

1. GSM 模型的性能较为优越；

2. 2D CNN 可以比部分时空网络效果好；

3. BERT 结构相对 2D CNN，性能更好，且 BERT resnet50 和 BERT bninception 性能较突出；

4. Slow 通道的迁移效果最差，不适合应用在 CEUS 中；

5. TSM 网络的迁移性能较差，这可能因为 Shift 模块的模式在 CEUS 上无法迁移利用；

6. I3D 的表现较为平庸；

7. C3D 对在更多数据时，性能提升明显，实现了 0.9 的正确率和 0.867 的 ICC 敏感性，在所有实验中变现最好的；

8. R(2+1)D resnet34 相比 resnet34+BERT，性能提升更明显, 基于 resnet 50 比 resnet 34 性能好，我们推断如果使用更深的 resnet 网络，性能会提升更高，因而是最有潜力的模型；

9. 提高网络的视频采样的图像数目能够普遍地提升分类性能；

10. 虽然原模型的训练基于密集采样，但均匀采样的迁移效果更好，因为数据包含了完整的造影剂增强和消退过程；

11. 数据清洗虽然去除了 21% 的图像，但平衡了 HCC 和 ICC 的数据量，直接使用原始数据时，有 3/4 的模型对 ICC 的敏感性变差，但在 C3D 上性能和敏感度都得到了提升，所以建议在具体使用对处理过和未处理数据集都测试；

12. non-local 结构虽然可以提升原问题的分类能力，但全局特征的关联不适

合迁移到 CEUS 中；

13. 在训练过程中，2D CNN 的时间和计算消耗最少，训练是，1000 次迭代在 5 分钟左右；2D+BERT 稍微慢一点，但比 GSM 和 TSM 快；R(2+1)D 的计算量较大；运行较慢；C3D，SlowFast, Slow 基于 3D 卷积，计算量最大，运行时间最长，1000 次迭代需要 20 分钟左右。

14. 对不同网络对所有病例分类结果可视化可以发现，个别病例在绝大部分网络中都无法正确识别，这种局限可能来自数据集，也可能是迁移学习无法获得区分力强的特征。

本研究是第一个系统研究视频模型在医学三维数据集中迁移能力的报告。

3.2　研究背景

3.2.1　迁移学习

迁移学习（Transfer Learning, TL）是机器学习中的一个研究问题，着重于存储在解决一个问题时获得的知识并将其应用于另一个但相关的问题。迁移学习可以分为三类：1）正迁移，即一种情况中的学习在另一种情况下促进学习时，例如，拉小提琴技巧有助于学习弹钢琴。数学知识有助于以更好的方式学习物理；2）负迁移，指学习一项任务会使另一项任务的学习更加困难时，例如，讲泰卢固语妨碍了马拉雅拉姆语的学习；3）零迁移，即学习一项活动既不能促进也不妨碍学习另一项任务。在医学影像中的深度学习研究领域中, 有很多迁移学习的成功应用，比如Samala 等 (2016) 使用深度卷积神经网络从乳房 X 线照片的迁移学习，开发用于数字乳房断层合成体积质量的计算机辅助检测系统，Huynh 等 (2016) 使用迁移学习对超声乳腺癌图像进行了表征，获得了强有力的结果，Apostolopoulos 和 Mpesiana (2020) 利用迁移技术从 X 射线图像中 Covid-19 进行自动检测，这些案例表明，网络的特征具有较好的泛化能力。从实践的角度来看，为学习新任务而重用或迁移先前学习的任务中的信息可能会显著提高强化学习代理的样本效率Karimpanal 和 Bouffanais (2018)。

最广为人知的迁移学习方法是在大数据集上从零开始训练网络，然后用需要分析的小数据集训用较小的学习率和较少的迭代次数微调（fine-tune), 该方法可以看做是把网络当成一个底层固定特征提取器，通过简单的调整高层特征的权重，获得小数据集的深层特征。与此同时，改进迁移学习效果的研究也在快速

发展，这些研究往往都基于大量实验，目前还没有重大的理论进展。在我看来，这些研究可以分为两类，1）改进训练策略，2）改进模型的泛化能力。

改进训练策略针对超参数的选择、数据集的优化和选择最优的迁移模块（网络每个子模块是固定，fine-tune 或者随机初始化）。结论具有借鉴意义的研究有：Dube 等 (2018b) 指出正确选择超参数和源数据集，可以提升 TL 的准确性。Dube 等 (2018b) 神经网络各层的学习率是重要的超参数，在 ImageNet22k 迁移到 Oxford Flowers 的任务中，选择适当选择学习率，可以将精度提高 127%；Ngiam 等 (2018) 指出更多的预训练数据并不一定有帮助，通过删除不相关的数据，改进预训练数据集不同数据的比重（给弱相关的数据低权重），在细粒度分类数据集上获得了卓越的性能；Chen 等 (2019) 对多个医学挑战中的数据集进行汇总，建立具有不同成像模态、目标器官和病理的 3DSeg-8 数据集，并通过建立名为 Med3D 的异构三维网络来共同训练多领域的 3DSeg-8，建立一系列预训练模型。相比直接从 Kinetics 数据集上的预训练模型，3DSeg-8z 上预训练的数据集在肺部分割、肝脏分割效果更好；Kornblith 等 (2018) 通过比较 12 个图像分类数据集上 16 个分类网络的性能得出，当网络用作固定功能提取器或微调时，ImageNet 精度与迁移精度之间有很强的相关性，此外，在两个小型的细粒度图像分类数据集上，ImageNet 中学到的要素无法很好地迁移。

改进模型的泛化能力的研究是研究什么样的网络结构可以更好地完成迁移任务，这些研究工作集中在迁移域学习（Domain Transfer, DL)，代表性的研究有：Huang 等 (2018) 设计了一个在结构上有多个组合的神经网络，能自动为输入在大的语义空间中选择最匹配的子特征空间。Dube 等 (2018a) 允许迁移学习在训练时和测试时的数据来自相似但不同的分布，通过自适应方法，寻找在源域上对主要学习任务具有辨别性且在不同迁移域中有较好表现的特征，以提高泛化能力。WEI 等 (2018) 提出了一种新颖的转移学称为” 学习转移”(Learning to Transfer) 的学习框架，通过元学习，使得网络在面对不同迁移问题时能自动确定那些特征需要保留，哪些需要训练。Wan 等 (2019) 指出虽然迁移学习通常可以通过更好的精度和更快的收敛来提升性能，但从不合适的网络中转移权重会伤害训练过程，并可能导致更低的精度，因为提出了一种正则化方法改进网络优化的方向使得网络部分参数得到较好的保留。

当然，在医学图像分析中也明确存在负迁移的现象，Raghu 等 (2019) 指出

ImageNet 预训练模型在两个大规模医学影像任务上，相比简单、轻量级的自定义模型，几乎没有性能上的提升。但研究也指出，迁移效果不好不是因为 ImageNet 上的特征不能用于医学数据上，而是由于标准模型的过度参数化，以至于在迁移问题上过度的拟合。本次研究也发现，网络的参数越大，越容易出现过度拟合，导致泛化效果越差。

3.2.2　数据集

本研究预训练模型使用的数据集有：

- ImageNet(Deng 等, 2009),2010 年发布的 2D 视觉对象识别软件研究的大型可视化数据库，包含 1000 个类别，每个类别包含 1000 张左右的高分辨率图像，数据集基于生活图片，比如“气球”、“草莓”，对推动深度学习的发展贡献巨大。
- Kinetics(Kay 等, 2017),2017 年发布，包含 400 个人类动作类，每个动作至少有 400 个视频片段，共计 306k 个片段，每个片段持续 10 秒左右，取自不同的 YouTube 视频。这些动作以人为中心，涵盖了广泛的类别，包括人与物的互动，如演奏乐器，以及人与人的互动，如握手。Kinetics 是视频分类任务领域首个大规模的高质量数据集，是视频识别方法评估的 base-line，也是很多数据集提供预训练模型。
- IG-Kinetics(Ghadiyaram 等, 2019),2019 年发布，目前规模最大的视频数据集, 数据未公开，包含 359 个类别，65M 个片段，数据来自社交媒体，没有经过人工筛选，标签较为嘈杂，通过弱监督学习初始化网络再 fine-tune 至 Kinetics 上实现了目前 Kinetics 最高正确率。
- Something-v1(Goyal 等, 2017b),2017 年发布的中等规模的视频片段数据集，共计 170 个类别，110k 个片段，致力于让机器学习模型对物理世界中发生的基本动作进行精细的理解，类别如“戳破一堆罐子”,“将电缆插入充电器”, 数据以每秒 12 帧的速度从原始视频中提取，并转化为 JPG 图像保存。
- Sport-1M(Mishra 等, 2013),2014 年发布，来自 YouTobe 的体育栏目视频，有 10k 个视频片段，分为 487 类，数据规模相对较小。

3.2.3　图像网络

由于视频网络是基于 2D 图像网络发展来，如 TSM 和 GSM 是在 2D 网络结构上加入时间轴上特征融合的模块实现，3D 网络也是简单的将 2D 网络的基本

元素扩展为 3D（2D CNN →3D CNN, 2D Pooling →3D Pooling), 学习视频网络需要以 2D 网络为基础。而且，视频网络 Base-line 也是基于对视频每一帧使用 2D CNN 识别，最后对每一帧的输出做简单的平均得到的。在 CEUS 中，对视频数据使用 2D+t 的应用也比较常见。本节简单介绍 2D 图像分类网络的发展，当然，由于新的改进不断涌现，大部分 SOTA(Sate of Art) 可能一夜之间被一项新的研究提升，此外改进训练策略或者数据集也是一个努力的方向，该部分必然是有局限的。

- 深度学习的复兴源于 2012 年提出的 AlexNet(Krizhevsky 等, 2017) 在 ImageNet 比赛中实现一骑绝尘的胜利，AlexNet 包含八层，前五层是卷积层，后三层是完全连接的层。它的激活函数——线性整流函数（Rectified Linear Unit, ReLU）使训练性能得到了改善，在 ImageNet 上最好的的 top5 Accuracy 为 71.194%。

- 2014 年之前的标准 CNN 结构是堆叠卷积层，通过最大池化，连接全连接层，但这类模型占用内存较大，容易过拟合。Inception(Szegedy 等, 2014) 网络为了减少网络尺寸，提出了 1 ×1 卷积，并在空间中使用不同规模的卷积核然后汇总，Inception-v1（GoogLeNet）使用 7 个 Inception 单元，平均池化和单层的全连接层;Inception-bn（Inception-v2)(Szegedy 等, 2015) 加入了 Batch Normlization(Ioffe 和 Szegedy, 2015), 去掉 5 ×5 卷积，改用两个 3 ×3 叠加的卷积, 提高了学习率和衰减系数（Weight Decay)，去除了 Dropout, 取消了 L2 正则化, 并指出 batch 越大优化效果越好；Inception-v3(Szegedy 等, 2015) 在 v2 基础上进一步缩小网络尺寸，指出特征图从输入到输出应该缓慢减小, 使用 stride 大小为 2 的池化层和卷积层，并且提升了网络的深度，将 n ×n 的卷积核分割为 n ×1 和 1 ×n 两层卷积；后来在Szegedy 等 (2016) 中提出的 Inception-v4 和 Inception-v3 较为相似，Inception-Resnet 加入残差网络的连接方式，Inception-ResNet 的精度和 Inception-v4 一致，在 ImageNet 上最好的的 top5 Accuracy 为 95.3%。

- VGG(Simonyan 和 Zisserman, 2014) 在 AleXNet 基础上做了改进，整个网络都使用了同样大小的 3 ×3 卷积核尺寸和 2 ×2 最大池化尺寸，引入 1 ×1 的卷积核，采用了 Multi-Scale 的方法来训练和预测，网络结构简洁，虽然在 ImagNet 中的排名较靠后，但在图像风格化领域展现出难以理解的优越性，在 ImageNet 上最好的的 top5 Accuracy 为 VGG19_ BN 的 92.66%。

- ResNet(He 等, 2016) 的提出是 CNN 图像史上的一件里程碑事件，最大深度

为 152 层，基本框架参考了 VGG19 网络，通过短路机制加入了残差单元，有效改善了深层网络的梯度消失和收敛慢的问题；此外，ResNet 直接使用 stride=2 的卷积做下采样，并且用全局平均层替换了全连接层, 为了保持网络不通层复杂度一致，当 feature map 大小降低一半时，feature map 的数量增加一倍。ResNetXt(Xie 等, 2016) 是 ResNet 的升级版，融合了 VGG 网络堆叠和 Inception 网络 “split-transform-merge” 思想，提出了 Group CNN，即将一个 CNN 层转换为多个尺度相同但通道数更少的 CNN，文章指出，该方法比增加宽度和网络深度更有效，在 ImageNet 上最好的的 top5 Accuracy 为 ResNeXt101_64 ×4d 创造的 94.252%。

- DenseNet(Huang 等, 2016) 吸取了 Resnet 的优点，对比于 ResNet 的 Residual Block，创新性地提出 Dense Block，在每一个 Dense Block 中，任何两层之间都有直接的连接，通过密集连接，缓解梯度消失问题，加强特征传播，鼓励特征复用，极大的减少了参数量，在 ImageNet 上最好的的 top5 Accuracy 为 DenseNet161 创造的 93.798%。

- SENet(Hu 等, 2017) 提出了一种全新的特征重标定策略，通过学习的方式来自动获取到每个特征通道的重要程度，然后依照这个重要程度去提升有用的特征并抑制对当前任务用处不大的特征。SENet154 在 ImageNet 上 top5 Accuracy 为 95.53%。

- SqueezeNet(Han 等, 2016) 是重要的压缩网络，用比 AlexNet 少 50 倍的参数达到了和 AlexNet 相同的精度, 大量使用 1 ×1 卷积核替换 3 ×3 卷积核，延迟下采样，没有全连接层，定义了 1 ×1 卷积核的 squeeze 层和混合使用 1 ×1 和 3 ×3 卷积核的 expand 层，在 ImageNet 上最好的的 top5 Accuracy 为 80.8%。

- PolyNet() 多项式的角度推导 block 结构，提出了更丰富的结构堆叠方式，文章指出虽然增加网络的深度和宽度能提升性能，但是其收益会很快变少，从结构多样性的角度出发优化模型，也可以提升性能，在 ImageNet 上最好的的 top5 Accuracy 为 95.75。

- DualPathNet(Chen 等, 2017) 在 ResNeXt 的基础上引入了 DenseNet 的核心内容，使得模型对特征的利用更加充分，在 ImageNet 上最好的的 top5 Accuracy 为 DualPathNet107 实现的 94.684%。

- NASNet(Zoph 等, 2017) 通过强化学习自动产生最好的网络结构，使用 Proximal Policy Optimization 方法优化，一次决策可以分解为输入两个并行卷积，

控制器决定选择哪些 Feature Map 作为输入以及使用哪些运算来计算输入的 Feature Map，再控制器决定如何合并这两个 Feature Map，MSNet 本质上是更为复杂 Inception，在 ImageNet 上最好的的 top5 Accuracy 为 NASNet-A-Large 实现的 96.163%。

- PNASNet(Liu 等, 2017) 是基于 NASNet 的改进网络自动生成方法，其训练时间为 NASNet 的 0.125，采用启发式搜索的策略:Sequential model-based optimization，在缩减的空间中进行搜索降低学习难度，增加 Agent 预测模型的精度，在 ImageNet 上的 top5 Accuracy 为 PNASNet-5-Large 实现的 96.182%。

3.2.4 视频网络

1. BERT（Bidirectional Encoder Representation from Transformers）(Devlin 等, 2018) 由 google 在 2018 年提出，提出时在 11 项自然语言处理任务中表现卓越，很快取代 LSTM 等改进的时序网络在自然语言处理中的地位。BERT 的网络结构基于Peters 等 (2017) 提出的 transformer 结构。transformer 集成了 self-attention(确定时序信息的关联，同时可以并行计算)、multi-attention（用更复杂的结构增强模型的表达能力）、position encoding（加入位置信息，以区别对待不同位置的输入），transformer 对输入和可能的输出都进行编码以确定二者的关联，BERT 只使用了 transformer 的输入编码结构，通过对 transformer 的堆叠实现层的概念，每一层都是 seq2seq，比如输入是 “Hello, World” 两个英文单词, 输出是 “你好，新世界” 两个中文单词，时序关系任然保留。输入序列中的每一个元素除了送入对应位置的 transformer, 还输入到该层其他的 transformer 中，这样 transformer 除了看到对应位置的输入编码，也可以看到全文，帮助消除歧义（apple 可能是苹果电脑，也可能是水果，需要根据语境判断）。本研究将 2D 网络编码后的视频理解为一个” 句子”，每个” 词” 为一张图像的卷积层输出特征，词的位置关系对应图像在视频序列中的排序。

在大规模视频数据没有出现前，CNN+LSTM 是一种常用的视频分析方法，但由于时间特征和空间特征融合得太晚，且 LSTM 的表达能力有限，在大数据集上表现一般，BERT 作为一种更新的时序网络结构，它的特征表达能力更强（结构更复杂），泛化效果更好（残差网络结构）, 所以本研究将 CNN 预训练模型 + 随机初始化的 BERT 结构纳入迁移学习的范畴。同时，考虑 2t+1D 网络更直接的原因在于实验发现 2D CNN 的性能和 3D CNN 能力相当，我们研究他们之间

的过度结构是希望了解 3D 网络的时空特征在 CEUS 迁移中是否是一个冗余的概念。

2. C3D(Convolutional 3D) 及其改进网络。3D CNN 的概念源于 C3D 网络 (Tran 等, 2014)，即使用 3 ×3 ×3 卷积核和 pooling，2D 网络的结构还以 AlexNet 为参考，只有五层卷积层接最大池化和三层全连接层，结构较为简单，输入大小为 3 ×16 ×112^2，代表 RGB 图像，clip 长度为 16，空间分辨率为 112 ×112 像素。但作者也表明，C3D 网络的参数量增加几乎没有提升网络的性能（当时还没有大规模数据集）。C3D 在 UCF-101 上正确率为 82.3%，在 Sport-1M 上正确率 85.4%。

Carreira 和 Zisserman (2017) 指出 3D CNN 看似更适合做视频处理，但它比 2D 有更多的参数，更难训练和泛化，因此提出了使用 2D 网络预训练数据集然后将网络参数扩展至 3D 维再训练，就是将 2D 卷积核的参数在新增的维度上膨胀（repeat）原有的参数来初始化。I3D 在 HMDB-51 上达到 74.8%，在 UCF-101 上达到 95.6%，在 Kinetics 上达到了 top1 72.1%。Wang 等 (2017) 在 I3D 的基础上，提出了 non-local 结构，基于计算机视觉中经典的非局部方法，将某一位置的响应计算为所有位置特征的加权和，以捕获长距离依赖信息，NL-I3D resnet101 在 Kinetics 上达到了 top1 77.7% 的精度，提升较为明显。Feichtenhofer 等 (2018) 在 I3D 的基础上提出了 SlowFast 网络，模型包括 1）低时间分辨率的慢速路径（Slow Pathway），以捕获空间语义，和 2）高时间帧率下分辨率的快速路径（Fast Pathway)，以捕获精细的时间分辨率的运动。快速路径可以通过降低其通道容量而变得非常轻量级，SlowFast 在视频中的动作分类和检测方面都表现出很强的性能，这些提升主要来自慢速路径，SlowFAst-NL-resnet101 在 Kinetics 上达到了 top1 79.8% 的正确率。

3. R(2+1)D 及相似的时间特征先分离再融合网络，该部分的发展基于传统的二维 CNN 计算成本低，但不能捕捉时间关系；基于 3D CNN 的方法可以获得良好的性能，但计算量很大，部署起来非常昂贵的背景。R(2+1)D 的提出源于Tran 等 (2017) 观察到应用于视频单个帧的 2D CNNs 在动作识别中仍然表现稳定，并且 3D ResNets 相对 3D inception 可以有效避免过拟合，因而将 3D Resnet 网络的卷积滤波器分解为独立的空间和时间组件，即将 t×d ×d 的卷积核转化为 1 ×d ×d 和 t ×1 ×1 的两步卷积操作，参数量缩小的同时，更容易优化, 在 Sport-1M 上 top5 正确率为 91.2%，Kinetctic top1 为 72.0%，Ghadiyaram 等 (2019) 在使用大

规模无监督预训练网络 R(2+1)D Resnet152 网络在 Kinetics 的 top1 为 79.9%，基于 Resnet34 达到 78.2%。

Lin 等 (2019b,a) 提出了一种通用有效的时移模块：Temporal Shift Module (TSM)。TSM 沿时间维度移动部分通道，从而促进相邻帧之间的信息交换，可以插入到 2D CNN 中, 以零计算和零参数的方式实现 3D CNN 的性能，即在 1 ×d ×d 操作后将空间编码的特征在时间维度向前或向后移动（shift)。Lin 等 (2019b) 可以部署在移动段 (Jetson Nano, Galaxy Note8), 实现低延迟的在线视频识别。在超级计算机 Summit 上使用 1,536 个 GPU 的进行规模化训练，将 Kinetics 数据集上的训练时间从 49 小时 55 分钟缩短到 14 分 13 秒，达到了惊人的速度和 74.1% 的精度。Sudhakaran 等 (2019) 指出 R(2+1)D 和 TSM 学习的都是结构化的内核，网络中的任何节点的连接不依赖输入数据，不能很好表现时间轴上复杂的关联，因此基于 TSM 设计了一个可学习的通道分割门控（splite gate）结构，使时间轴上的特征可以选择 shift 的类型，并且使用残差网络，防止过拟合。GSM 在 Something-v1 上实现了 top1 55.16% 的突破性进步。

3.3 实验设计

3.3.1 数据预处理

1. DICOM 导出为图像序列：使用RadiAnt DICOM VIEWER 4.6.5 软件将从机器导出的 DICOM 视频数据转为图片序列，图片命名为” 肿瘤类型_视频编号_图片在序列中的排序数.bmp”，同一病例图像序列存放在一个文件夹中。

2. 添加时间标签：由于增强超声视频的时间采样频率在超声机器上可以调整，原始的 DICOM 数据可能没有记录确切的帧率，或者在视频导出过程中的处理不当导致帧率记录错误，图片的在序列中的录制时间不能通过简单的方程 (3.1):

$$\frac{frame\_index}{total\_frames} \times (end\_time - begin\_time) + begin\_time \quad \ldots (3.1)$$

由于 CEUS 录制时图像上的固定位置会显示录制的时间，如开始录制后 12s 显示为 “00:12”，一般，相同机器导出相同大小的视频时间文本位置固定，我们对每一个视频的第一张图像的时间文本手工定位后使用 Matlab 2019a 软件中的Optical Character Recognition (OCR) 工具包自动识别整个序列每张图像的时间标签，并将图像的命名方式改为 “肿瘤类型_视频编号_OCR 时间_图片该时刻的排序数”。

3. 去除噪声数据：由于计时一般在造影剂打入后开始，最初的十几秒，造

影剂还没有到达肝脏，此时图像中只有背景噪声，而每个人开始增强的时间也有较大差异。此外，部分视频在 2 分钟之前病灶就消退完全，在此之后的数据也基本为背景造影，因为医生在画病灶区域时确定了开始增强时间和消退时间，通过检索图片的命名，将每个视频增强前和消退后的图片删去。该操作将 314 120 张图像缩减为 289 233 张。

4. 平衡数据分布：虽然 HCC 和 ICC 的入组病例数一致，但由于 HCC 的消退时间较晚，ICC 一般早于两分钟消退，实际医生在录制的时候，当造影剂消退了便会停止录制，整理后的数据中，HCC 的图片数大概是 ICC 的 1.3 倍，同时由于 CEUS 的帧频较高且可以调制，部分病例的图片数达到了 3000 张，而时间较短的视频只有 500 多张，为了平衡 HCC 和 ICC 的数据量和各个病例所占的比重，我们将每秒图片张数大于 7 的序列进行随机采样，只保留每秒 7 张图像的帧率，然后计算平均每个视频的帧数，对大于平均值的视频进行随机采样至平均张数，最后，我们对整个数据集进行随机采样使得 HCC 和 ICC 的图像数一致。通过这个操作，我们分别删减了，最终 HCC 的图片数为，ICC 的图片数为。该操作将 289 233 张图像缩减为 248 262 张。

5. 数据增强：视频的增强方法使用现将图像由 RGB 转化为灰度图像，这是因为超声造影是伪彩图像，转化为灰度图像，一方面可以不同机器彩色化显示时不同的 Colormap 使图像外观有差异的影响，另一方面可以降低内存加快图像处理速度。本实验没有直接将一个完整视频直接送入网络，而是在时间轴上降采样得到多个 Clip，每个 Clip 独立送入网络分类，再将一个视频的多个 Clip 分类结果取平均。训练集的 Clip 或者图片操作流程：在空间轴上 Resize 到比网络标准输入大小多 0-32 个像素的尺寸 → 通过 Color Jittering(Devries 和 Taylor, 2017) 对一个 Clip 或者单张图像的亮度 ±0.03、饱和度 ±0.03、对比度 ±0.03 → 发生概率为 0.5 的左右镜像操作 → 以 0.5 的概率发生的，对空间轴上某个位置随机的大小为 5000 像素的矩形区域置的像素随机生成 (Random Black)(Zhong 等, 2017) → 对一个 Clip 或者单张图像 Random Crop 成网络标准输入的大小，图像的均一化参数与使用网络的给定值一致。测试集中的数据操作为：Resize 到网络标准输入大 16 个像素的尺寸，再通过 Center Crop 和 Normalization 转化成网络标准输入。

6. 固定测试数据：虽然采用了交叉验证，但考虑随机采样生成测试样本在性能比较时，每次测试的数据会有差异，因而对每个交叉验证组的测试集固定

测试数据，一次性地随机在每个测试序列中按采样方法生成 300 个 clip 或者 500 张图片，通过灰度化，Resize 到固定大小后使用 Centercrop 去除上下左右各 8 个像素，通过减均值去方差操作后，复制到三个 RGB 通道并保存数据。

3.3.2 网络训练

所有网络训练实验均使用的 SGD 优化器，损失函数为交叉熵损失，fine-tune 过程不对任何网络结构做固定参数的操作，采用固定的学习率。由于研究的是 fine-tune 的效果，实验的学习率设置得较低，保证网络训练过程不出现发散，损失函数不断地收敛。需要注意的是，当将学习率设置较大，迭代次数设置较高，迁移学习将转变为将源数据训练的参数作为网络的初始化值，然后在目标数据集上重新学习特征，这时我们研究的问题变为各个网络能否拟合 CEUS 数据，这与本研究的初衷相违背。

实验使用的服务器配置了 4 块 Intel(R) Xeon(R) Gold 6130 CPU 和 4 块 NVIDIA TITAN Xp GPU（12 GB 显存），基于 Anaconda 环境，使用 Python 3.6.0，显卡驱动版本为 NVIDIA-SMI 440.82，CUDA 版本为 10.2，深度学习框架为 PyTorch1.4.0。具体使用的预训练模型如下：

1. 2D CNN 网络的定义的参数来自 Githb 项目Pretrained models for Pytorch，网络的预训练数据集为 ImageNet。2D CNN 方法在训练时将视频中的每一个图像作为一个独立输入，标签为该图片是否来自 HCC 或者 ICC，测试时将每个视频保存的 500 张图像一次送入网络，网络每次对单张图像预测，得到一个长度为 500 的值为 0 或者 1 的向量，当向量的和大于 250 时，判定视频为 ICC, 否则为 HCC。由于本次研究网络视频的基本框架建立在 ResNet34、ResNet50、ResNet101、BnInception、Mobilenet-v2，我们专门测试了对应 2D 网络的性能，对比改进的模块是否在 CEUS 迁移学习中表现更好。

2. 2D CNN_BERT 网络使用的前端模型和参数与 2D CNN 网络一致，使用 2D CNN 提取图像特征，再在时间轴上将这些特征打包为一个数据块送入 BERT 模块中完成分类任务。本研究使用的 BERT 参考官方代码更改为 pytorch 模型，参考 Alexnet 在卷积层后接 3 个全连接层的方式，实验设定 BERT 层数为 3，其他超参数通过在 AlexNet2D 网络上测试调整，最终确定为输入向量长度 1000，中间层向量长度为 1280，Dropout 比例为 0.2，multi-head attention 个数为 5。

3. C3D 网络参数来自个人项目c3d-pytorch。

4. I3D 和 SlowFast,Slow 网络均来自 Facebook 开源项目SLowFast。
5. R2plus1d 来自 MicroSoft 计算机视觉模型开源库VZM。
6. TSM 来自论文作者 Github 开源项目temporal-shift-module。
7. GSM 来自论文作者 Github 开源项目GSM。

为了保障迁移效果不受各个项目特殊的操作影响（如 warm_up，数据增益），我们只使用这些项目的网络结构定义文件和预训练模型，实验的数据接口等其他部分由我们自行编写，减少因训练 tricks 的不同带来的差异。研究涉及的超参数有学习率，SGD 的动量, 衰减率, 迭代次数，其中学习率, 衰减率和动量采用 grid-search 确定，学习率的搜索范围为：0.001，0.0003，0.0001，0.00003；衰减率的搜索范围为：0.005，0.001，0.0001，0；动量搜索范围为：0.9，0.99，迭代次数的设置为了节约时间，设置最大值后一次性迭代完成，每隔固定间隔，执行一次测试，并且保存模型，最后选择最优的模型为最终性能。由于 2D CNN 网络和 2D CNN+BERT 的拟合速度快，为了节约时间，最大迭代次数设置分别为 3000 次和 6000 次，每次测试的间隔设置为 1000，其他模型的最大迭代次数设置为 12000，每次测试的间隔设置为 2000。BatchSize 的大小由网络的内存占用量决定，设置为显卡一次计算能允许的最大值。本实验在时间轴上图像采样的方法为 uniform sample，比如网络输入是 8 张图像，则将视频等分为 8 段，使用采样器随机在视频上找到一张图像，计算图像属于视频中的哪一段，在计算其在该段中的比例，再在其他分段中找到和它相同比例的图片，汇总成一个 clip。在性能测试中，除了统计单个视频基于 300 个 clip 的分类结果，我们同时也对单次输入 clip 的正确率和 AUC 进行计算。

3.4　实验结果

3.4.1　基本模型

一次系统地参数遍历后对最优模型汇总，得到表 3.1和图 3.1，这些模型都基于 $3 \times 8 \times 224^2$ 的输入尺寸，其中表格中预训练模型的正确率为在原来的数据集上的 top1 正确率，模型参数量的单位为 M(1 million)，迭代次数的单位为 k(1 thousound)，* 号表示没有对应的预训练模型或者参数未知，Slowfast 是双通道模型，输入视频帧数用 “Slowframes（Fastframes）” 表示。图 3.1为 Heatmap 图，横坐标对应不同的分类模型，纵坐标代表视频编号，每个位置的颜色块代表该模型

对该视频的分类结果（[0,1] 之间的网络得分，1 为 ICC，0 为 HCC），我们可视化该结果是为了了解不同模型的错误分类结果是否有关联（是否有病例，所有模型都错误判断）图像的上半部分为 ground truth 是 HCC 的视频，下半段是 ICC 视频，理想情况下，图像的上半段颜色值都 ≤0.5，图像的下半部分都 ≥0.5, 这是代表所有病例都被正确识别。

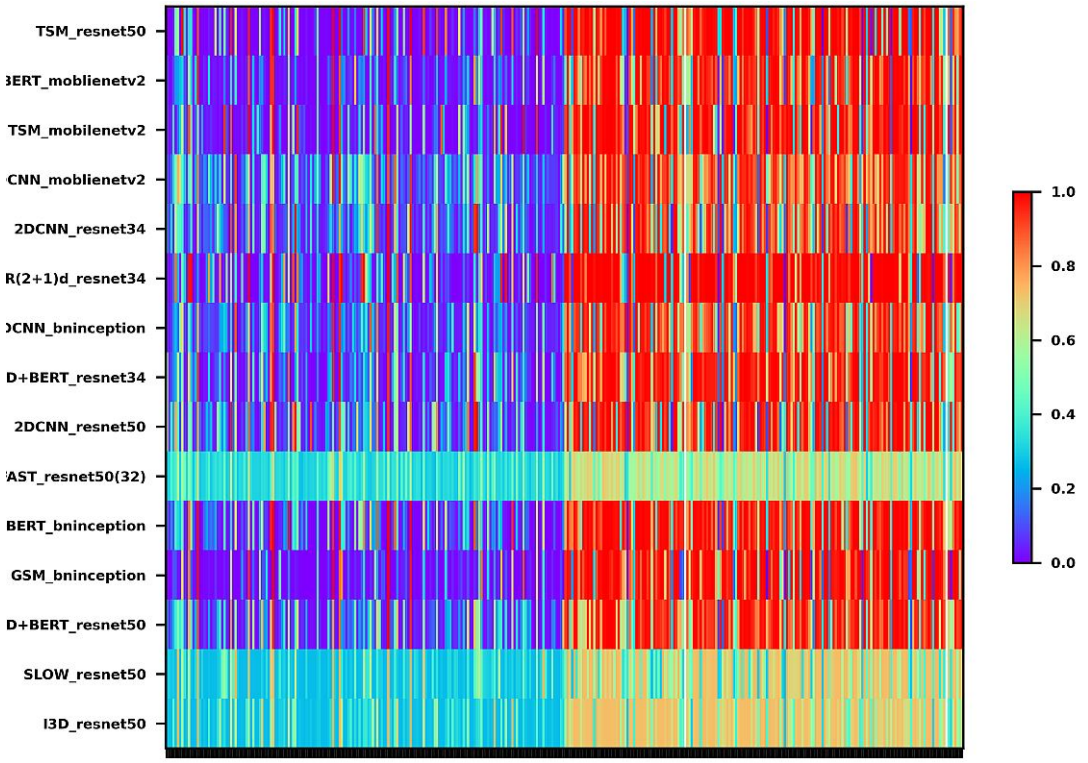

图 3.1 网络测试结果 Heatmap 图，水平方向的刻度代表病例编号，竖直方向为网络结构名，每个位置的方块代表该网络对于此病例的是 ICC 的评分，图像的左半部分是 HCC 病例，右半部分是 ICC 病例。

Figure 3.1 Heatmap of network test results, the horizontal scale represents the case number, the vertical direction is the network structure name, each position of the box represents score that the network think the vedio is ICC. The left half of the image is HCC cases, the right half of the image is ICC cases.

通过图表，可以发现表 3.1可以发现，2D CNN 的性能和复杂的时序网络相比，性能不算很差，这和 Kenetics 分类任务中 2D CNN 的性能也不是很查的现象一致，这说明了 HCC 和 ICC 的外观结构属性在较多的造影图像中可以区分出来，而不必知道时间信息。2D CNN 再加入 BERT 以后，除了 bninception 模型外都出现了普遍的提升，尤其重要的是，2D moblienet+BERT 在所有网络中正确率

表 3.1 固定输入 8×224^2 时模型迁移性能

Table 3.1 model transfer learning performance when the input size is 8×224^2

迁移模型					训练参数				测试结果						
模型	骨架	数据集	输入大小	正确率 (%)	参数量（M)	学习率	衰减率	动量	迭代次数 (k)	正确率	AUC	敏感度	特异性	1-clip 正确率	1-clip AUC
2D CNN	resnet 34	Image Net	1 ×224^2	73.55	21.5	0.0003	0.005	0.99	2	0.8352	0.9077	0.7908	0.865	0.8146	0.848
	resnet 50	Image Net	1 ×224^2	76.02	25.3	0.001	0.005	0.9	1	0.8575	0.9047	0.7995	0.8673	0.8427	0.8671
	bnin -ception	Image Net	1 ×224^2	73.52	<25	0.001	0.005	0.9	1	0.8799	0.9452	0.8295	0.9062	0.8989	0.8649
	mobile - netv2	Image Net	1 ×224^2	71.8	3.47	0.0003	0.005	0.99	1	0.8687	0.9422	0.8206	0.8992	0.8933	0.8503
BERT	resnet 34	Image Net	8 ×224^2	73.55	21.5	0.0001	0	0.99	6	0.8659	0.9446	0.8372	0.9158	0.882	0.8533
	resnet 50	Image Net	8 ×224^2	76.02	25.3	0.0003	0	0.99	3	0.8743	0.9446	0.8378	0.9133	0.9045	0.8519
	bnin -ception	Image Net	8 ×224^2	73.52	<25	0.0003	0	0.99	2	0.8687	0.9365	0.834	0.9081	0.8652	0.8701
	mobilenet v2	Image Net	8 ×224^2	71.8	3.47	0.001	0	0.99	4	0.8799	0.9422	0.8386	0.9134	0.8483	0.9042
C3D	*	Sport-1T	8 ×224^2	*	*	*	*	*	*	*	*	*	*	*	*
I3D	resnet 50	Kinetics	8 ×224^2	73.5	35.3	0.0001	0	0.99	6	0.8687	0.9263	0.8292	0.8979	0.8708	0.8659
Slow fast	resnet 50	Kinetics	8(32) ×224^2	77	32.4 +0.53	0.0001	0.0001	0.9	2	0.8575	0.9104	0.7898	0.8438	0.8539	0.8588
Slow	resnet 50	Kinetics	8 ×224^2	74.8	32.4	0.0001	0.001	0.9	10	0.8184	0.8952	0.8001	0.8744	0.7528	0.8645
R(2 +1)D	resnet 34	ig65	8 ×112^2	*	33.3	0.0001	0.001	0.99	4	0.8771	0.9324	0.8546	0.9245	0.9157	0.849
TSM	resnet 50	Kinetics	8 ×224^2	74.1	25.3	0.001	0	0.9	2	0.8408	0.9182	0.8318	0.9065	0.8427	0.838
	mobile - netv2	Kinetics	8 ×224^2	69.5	3.47	0.001	0	0.9	6	0.8687	0.9253	0.8491	0.9152	0.8876	0.8541
GSM	bnin -ception	Some thing-v1	8 ×224^2	49.01	10.5	0.0003	0	0.9	8	0.8631	0.9358	0.8274	0.9045	0.8427	0.8772

最高，其次是 2D Bninception, Bninception 的优秀表现可能来自没有残差结构，网络的 fine-tune 逐层进行，底层的特征得到较好的保留。BERT 的性能表明了时序关系可以通过 2D CNN 转化到高级的语义空间，当成词向量理解成一个 “句子”，做一个类比，医生可以用语言描述造影视频每隔每一段时间造影图像的外观，然后让另一个医生判断这是什么疾病。BERT 的优秀表现，很可能跟造影视频的特定肿瘤的增强模式在时间轴上有规律可循有关。不同于自然场景的视频截取，一个动作可以发生在前几秒或者后几秒，而且事件的切换也比较随机，如两个人见面可以先握手再拥抱，也可以先拥抱再握手，但造影视频的达峰时间一定在开始增强后 30 秒以内，增强一定发生在衰减前。

由于 C3D 的输入尺寸要求为 16×112^2，因而不包含在本对比中，它在性能在表 3.2中。

I3D 和 SlowFast 以及 Slow 这三个网络的性能并不突出，其中 Slow 是所有模型中正确率最低的，对比 SlowFast 模型，我们认为这可能和 SLow 网络网络增加了时间轴上的参数但弱化时间的联系以强调空间有关，这个结构在功能上近似 2D CNN，但又有更多的参数。

R(2+1)D resnet34 网络是在基于视频数据预训练的网络中性能最好，这可能跟它的预训练数据集规模庞大有关，也可能和它自身的设计简洁，从局部逐渐向上的特征提取方式有关。对比 resnet34 网络，性能的提升比 2D+BERT resnet34 好。

TSM resnet50 的性能较 2D resnet50 差，与此同时 TSM mobilenetv2 的性能比 2D mobilenet 差，由于 shift 操作是零参数，对应网络的参数量一致，但新加入的功能反而起到了负面的影响，这可能跟 shift 操作在原问题上的模式不适合迁移到 CEUS 上有关。GSM 基于 bninception，虽然使用了 shift，但通过门控操作可以适应输入数据的做出改变，且通过残差网络，可以跳过不必要的 shift 操作，性能较好且计算量小。

3.4.2 采样帧数

在视频分类任务中，一般提高时间轴上的图像个数，能够提升网络的性能，不同与 8 的采样结果模型测试结果如表 3.2，对比表 3.1，可以发现增加输入量使得所有网络的性能都得到提升，C3D 网络表现较为中庸，当降低 Slow 通道的输入图像数目，SlowFast 和 Slow 网络的性能都下降了，R(2+1)D 和 GSM 的正确率

在所有模型中最高，但 GSM 对 ICC 的敏感度更好。

表 3.2 不同采样帧数下模型迁移性能表

Table 3.2 model transfer learning performance under different sample frames

迁移模型					训练参数				测试结果						
模型	骨架	数据集	输入大小	正确率 (%)	参数量（M)	学习率	衰减率	动量	迭代次数 (k)	正确率	AUC	敏感度	特异性	1-clip 正确率	1-clip AUC
BERT	**resnet 34**	Image Net	32 ×112²	76.02	25.3	0.0003	0	0.99	1	0.8771	0.9462	0.8426	0.9171	0.8876	0.8681
	resnet 50	Image Net	16 ×224²	76.02	25.3	0.0003	0	0.99	5	0.885	0.951	0.852	0.926	0.893	0.878
	bnin -ception	Image Net	16 ×224²	73.52	<25	0.0003	0	0.99	3	0.863	0.928	0.826	0.893	0.882	0.849
	mobile - netv2	Image Net	16 ×224²	71.8	3.47	0.001	0.0001	0.99	6	0.883	0.944	0.850	0.914	0.876	0.886
C3D	*	Sport - 1T	16 ×112²	84.4	79	0.0001	0.0001	0.9	12	0.867	0.942	0.789	0.921	0.871	0.879
Slow fast	**resnet 50**	Kinetics	4(32) ×224²	75.6	32.4 +0.53	0.0001	0.0001	0.9	10	0.8631	0.9276	0.7956	0.8507	0.8764	0.8525
Slow	**resnet 50**	Kinetics	4 ×224²	72.7	32.4	0.0001	0.001	0.9	4	0.8073	0.8811	0.7908	0.8576	0.8146	0.8011
R (2+1) D	**resnet 34**	ig65	32 ×112²	*	33.3	0.0001	0.001	0.99	10	0.8855	0.935	0.8288	0.877	0.9382	0.8477
TSM	**resnet 50**	Kinetics	16 ×224²	72.6	25.3	0.001	0	0.9	8	0.8743	0.9391	0.8568	0.9345	0.8539	0.8889
GSM	**bnin -ception**	Some thing -v1	16 ×224²	50.63	10.5	0.0003	0	0.9	8	0.8855	0.946	0.873	0.9374	0.8539	0.9102

3.4.3 密集采样

实际上，在 Kinetics 等数据的模型训练往往基于密集采样（dense sample），即每隔几张图像采集一张，或每隔一秒采集一张的方式，这样可以识别视频中精细时间尺度下的变化，对于发生速度快速的动作尤其有必要，但这个方法生成的数据在时间轴上跨度较小，由于 CEUS 的变化持续时间久，不是 Kinetics 中十几秒的短视频，理论上，均匀采样涵盖的信息更全，效果会更好。我们对 2D CNN_bninception，2D+BERT_resnet50_16，C3D_16,GSM_bninception_16 这四个在所属子类中性能较好的网络在保持最优参数不变的情况下，使用密集采用，结果如表 3.3：

表 3.3 密集采样时模型迁移性能表

Table 3.3 model transfer learning performance under dense sample

迁移模型					训练参数				测试结果						
模型	骨架	数据集	输入大小	正确率(%)	参数量（M)	学习率	衰减率	动量	迭代次数 (k)	正确率	AUC	敏感度	特异性	1-clip 正确率	1-clip AUC
2D CNN	bnin -ception	Image Net	224^2	73.52	<25	0.0003	0	0.99	2	0.832	0.885	0.779	0.847	0.837	0.828
BERT	resnet 50	Image Net	16 ×224^2	76.02	25.3	0.0003	0	0.99	3	0.858	0.921	0.816	0.880	0.876	0.843
C3D	*	Sport -1T	16 ×112^2	84.4	79	0.0001	0	0.9	12	0.902	0.958	0.853	0.924	0.899	0.904
GSM	bnin -ception	Some thing -v1	16 ×224^2	50.63	10.5	0.0003	0	0.9	8	0.866	0.938	0.838	0.910	0.876	0.857

通过表 3.3中的数据和和表 3.2中对应模型的性能，我们发现正确率都出现了明显的下降，1 clip 的真确率也都下降了，且 ICC 的敏感度变差。3.3表明短时间的造影 Clip 不能提供较多的时间轴上动态变化的信息帮助分类器做出正确的决定，均匀采样因为提供更完整的造影变化过程，迁移时，即便原模型是密集采样，仍然效果更好。

3.4.4　数据清洗

在实验中我们为了平衡 HCC 和 ICC 的数据量以及每个病例图片数量，缩减了 21% 的图片，为了探讨这样的操作对性能的影响，我们对 2D CNN_bninception，2D+BERT_resnet50_16，C3D_16,GSM_bninception_16 这四个在所属子类中性能较好的网络在保持除迭代次数外其他参数不变的情况下，使用全部的数据训练网络，结果如表 3.4:

表 3.4 完整训练集上模型迁移性能表

Table 3.4 model transfer learning performance on the full training set

迁移模型					训练参数				测试结果						
模型	骨架	数据集	输入大小	正确率 (%)	参数量（M)	学习率	衰减率	动量	迭代次数 (k)	正确率	AUC	敏感度	特异性	1-clip 正确率	1-clip AUC
2D CNN	**bnin -ception**	Image Net	224^2	73.52	<25	0.0003	0	0.99	2	0.832	0.885	0.779	0.847	0.837	0.828
BERT	**resnet 50**	Image Net	16 $\times 224^2$	76.02	25.3	0.0003	0	0.99	3	0.858	0.921	0.816	0.880	0.876	0.843
C3D	*	Sport - 1T	16 $\times 112^2$	84.4	79	0.0001	0.0001	0.9	12	0.902	0.958	0.853	0.924	0.899	0.904
GSM	**bnin -ception**	Some thing -v1	16 $\times 224^2$	50.63	10.5	0	0.0003	0.9	8	0.877	0.949	0.867	0.941	0.871	0.881

通过对比表 3.4和表 3.2中对应模型的性能，我们发现在 2D CNN_bninception，2D+BERT_resnet50_16,GSM_bninception_16 都出现正确率的下降，但在 C3D_16 中，正确率和敏感度都提升了，其他模型的 ICC 敏感度都下降，这意味着更少的 ICC 被正确判断，更严重的医疗事故将发生，这个现象出现的原因很可能是数据集中有更多的 HCC 图片，因而出现了较高的 Bias，而 C3D 性能的提升很可能跟它参数量大，需要更多数据拟合有关。

3.4.5　non-local 全局特征

在视频分类任务中，non-local 作为一个有效的模块，能够学习输入的全局关联，在 I3D 中可以有效提升性能，我们研究了 TSM_NLN_resnet50_8 和 I3D_-

NLN_resnet50_8，其中超参数的选择除了迭代次数外，与各自未加 NLN 的模型参数一致，结果如表 3.5:

表 3.5 non-local 模型迁移性能表

Table 3.5 non-local model transfer learning performance table

迁移模型					训练参数				测试结果						
模型	骨架	数据集	输入大小	正确率(%)	参数量（M)	学习率	衰减率	动量	迭代次数 (k)	正确率	AUC	敏感度	特异性	1-clip 正确率	1-clip AUC
I3D	nln resnet 50	Kinetics	0.74	35.5	8×224^2	0.0001	0.001	0.9	6	0.855	0.932	0.832	0.900	0.854	0.854
TSN	nln resnet 50	Kinetics	0.756	25.3	8×224^2	0.0001	0.001	0.9	10	0.858	0.925	0.828	0.901	0.837	0.871

通过对比表 3.1中的 I3D 模型和 TSM mobilev2，我们发现网络的性能都出现了下降。由于迁移学习是基于卷积特征能够将图像基本模块（边，角）提取出来，这些基本元素可以很好地利用在其他数据集上实现的。non-local 模块相比卷积算子，它的计算结果涉及整个输入，很容易猜想到，由于 CEUS 的视频内容不同与 Kinetics 数据集，他们的特征连接方式必然是差异较大的，正如搭建积木的材料是一样的，但最终搭建出来的东西完全不同，则自搭建方法必然是的不同的类比，使用 Non-local 这样的全局连接方法是不利于 CEUS 迁移学习的。