计算机视觉实验报告

项目名称：　ARDS 肺部轮廓分割　

小组成员及工作（即报告顺序）：

58120309 王玟雯、58120103 林佩君（组长）

58120217 俞家琛、58120127 蔡英豪

项目背景：

ARDS（急性呼吸窘迫综合征）是由肺内外原因引起的，以顽固性低氧血症为显著特征的临床综合征，因高病死率而倍受关注。在 2019 年席卷全球的新冠疫情中，ARDS 作为新冠重症病人在死亡之前的最后疾病状态，其病情进展迅速而凶险，需要高效而快捷的诊疗方案。CT（电子计算机断层扫描）影像具有能反映出 ARDS 患者肺部生理信息的能力，目前临床上主要借助 CT 影像来诊断患者病情并给予救助。然而 ARDS 病人的肺部往往呈现较多病变，如：塌陷、实变、积液、纤维化、毛玻璃影等，但目前存在的肺部轮廓自动标注工具难以完整地分割 ARDS 病人肺部轮廓，在实病变区域分割的准确率较低。在自动分割后，临床工作者往往需要手动修改肺部的轮廓边界。该工作通常需要 3-4 小时，严重影响临床的工作效率。

然而，在实际的实验过程中，我们发现相比于正常人的肺部，ARDS 病人的肺部 往往呈现较多病变，如:塌陷、实变、积液、纤维化、毛玻璃影等，传统的医学分割 U-net 网络在 ARDS 病人的肺部实病变区域分割的准确率较低，难以满足临床需求，更加高效并且高质量的肺部分割方式成为当下所需。在这样的可能需求下，我们小组展开了肺部分割相关方面的调查。

一、前期准备：数据集收集并实现 F3Net 医学分割领域应用迁移

1.1 F3Net 模型引入：

对这样的病变肺部进行分割，除了参考传统的医学影像分割任务，一定程度上也类似于伪装物体检测任务，即目标物体被不同程度的遮盖。基于这个思考，我们参考文献，深入了解伪装物体检测领域。

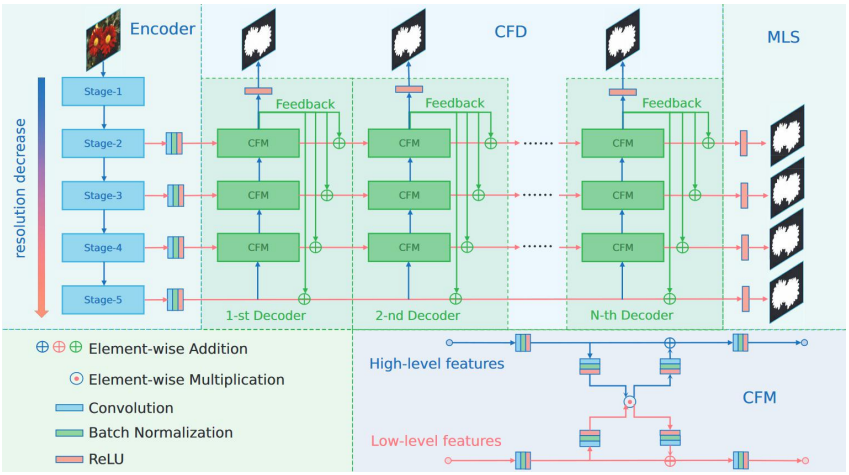


图 1-1: F3N 网络结构概览

现阶段伪装物体检测任务 SOTA 模型为 F3Net。我们参照 F3Net 模型，在 Unet 上引入了多层解码器结构，让不同维度的特征融合，用多级特征融合模块，融合高分辨率与低分辨率的图像特征，更大程度地减少了精度损失。在之后我们会更进一步对模型做出改进，此处我们先对于参照 F3Net 的模型进行应用迁移，以方便之后和我们设计的 FMEDNet 方法进行比较。

1.2 公共肺部影像数据集搜集

为了验证参照 F3N 网络进行改进的网络模型性能，在手头已有 CT 数据集的基础上，我们需要更多有标签的肺部影像 CT 数据进行训练。

数据集收集工作的难点在于：

1. 受限数据集的专业性，数据集查找没有头绪
2. 对专业医学影像数据格式（如.dicom, .mhd, .raw 等）的不熟悉
3. 模型需要数组形式的输入，不知道如何进行上述专业医学影像数据格式的转换
4. 数据清洗、数据和标签进行对齐、封装成可以直接使用的数据集类

我们尝试了以下数据集：

SIIM-ACR Pneumothorax Segmentation | Kaggle SIIM-ACR 肺部图像分割数据集

Download - Grand Challenge (grand-challenge.org) luna16 数据集

https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD TCGA 肺部数据集

最终我们选择了 luna16 数据集，其全称为 Lung Nodule Analysis 16。选择原因：

1. 它是 16 年推出的一个肺部结节检测数据集，旨在作为评估各种计算机辅助检测系统任务的 benchmark。专业性强，广泛应用于计算机分析领域。
2. 数据效果好。它的数据来源于一个更大的数据集 LIDC-IDRI，是将切片厚度（slice thickness）大于 3mm 的 CT 去除，同时将切片 space 不一致以及缺失部分切片的 CT 也去除，最后筛选出了更高质量的 888 张 CT。
3. 数据完整，标签质量高。标签 mask 无残缺，均和样本对齐。

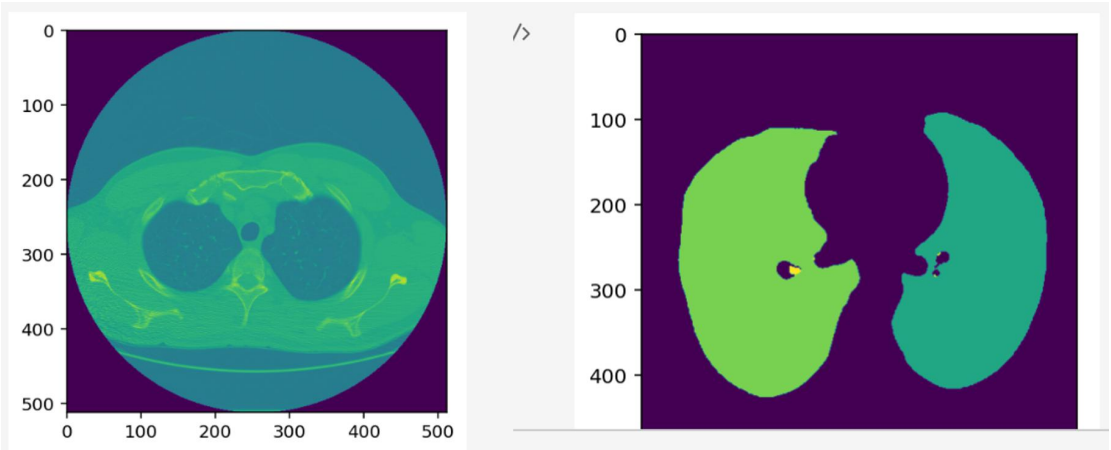


图 1-2: LUNA16 数据集样本展示

1.3 数据预处理

原始 CT 影像数据格式为 .mhd 后缀的医学影像文件。而在输入网络前，我们需要将每份原始影像数据都转化为.csv 数据格式文件。

为此，我们首先进行数据清洗与确认，确保标签没有缺漏并且和数据对齐。接着调用 SimpleITK 包，编写相应 ct2csv 代码文件大批量转化原始数据至 csv。Data 和 label 的结果的命名中各自标明 case_id 方便后续使用。数据转化实验在一块 1080Ti 上运行 3h 实现，转化近 4G 的公共医学影像数据以及其标签至 csv 文件。

1.4 F3N 模型迁移结果展示

利用获取的公共医学影像数据集，我们在根据 F3Net 改进的网络上进行模型迁移，训练其在医学影像分割领域的效果。我们使用 Dice Coefficient，IoU 等指标进行衡量模型

作为图像分割评价指标之一，Dice Coefficient 是用于评估两个样本的相似性的统计量，本质上是衡量两个样本的重叠部分。Intersection-Over-Union (IoU)，也称为 Jaccard 指数，是语义分割中最常用的指标之一。IoU 是预测分割和标签之间的重叠区域除以预测分割和标签之间的联合区域（两者的交集/两者的并集）。而 Mean IoU 就是分别对每个类计算（真实标签和预测结果的交并比）IoU，然后再对所有类别的 IoU 求均值。

迁移结果如下表所示

Model	Dice Coefficient	Global correct	Average row correct			IoU			Mean IoU
Unet	0.854	99.6	99.8	95.4	95.2	99.6	92.6	93.6	95.3
F3Net	0.891	99.6	99.9	95.4	94.8	99.6	93.3	93.0	95.3

表 1-1: 迁移到 F3N 模型后在测试集上的对比表现，各项指标都是越高代表约好的得分

二、FMED-Net 的整体框架提出

1.1 F3N 模型局限

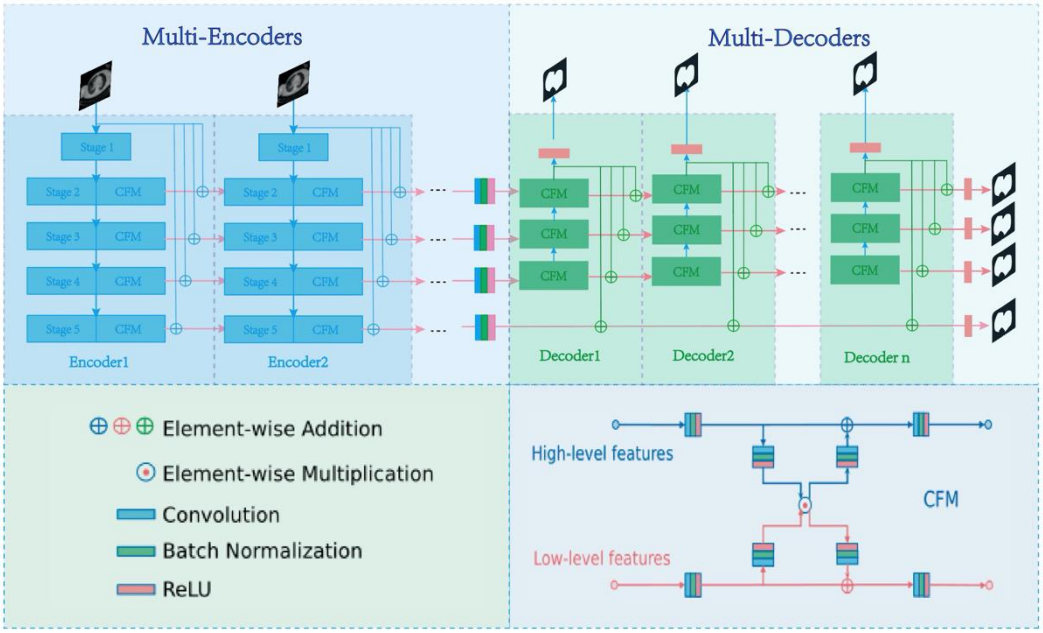
在 F3Net 网络结构中，模型通过多层解码器的级联，一定程度上弥补了 CT 影像在编码器中下采样时精度的损失。然而这种“post-remedy”的处理方法依然无法在本

质上解决精度损失的问题。我们发现，在肺部影像的边沿细节区域，F3N 网络分割效果较差。

我们需要在产生精度损失的下采取步骤中作出改进，在最底层最大程度减少进度损失。

1.2 从 F3N 模型局限出发，设计 FMED-Net 网络

基于这个动机，我们提出了 FMEDnet(Fusion mulit encoders-decoders net)，采用 fusion 的多层编 码器、多层解码器级联的结构。在 F3Net 多层解码器的基础上，对称地引入多层编码器级联，从根源处最大程度降低精度损失。完整 FMEDnet 网络结构如下图。



总模型由 Multi-Encoders 编码器模块和 Multi-Decoders 解码器模块组成。在编码器模块中，我们创新性地采用了多级微编码器子模块级联的方式。在每个微编码器模块上，我们使用了“refine”策略（在第四部分进行详解及拓展），利用高精度、低语意细节来弥补低精度、富语义层级的精度损失。每个微编码器之间相互级联，使用 cross feature module(CFM)融合不同层级编码器的输出结果，通过迭代的方式进一步减少精度损失。

1.3 FMED-Net 实验结果及实际效果分析

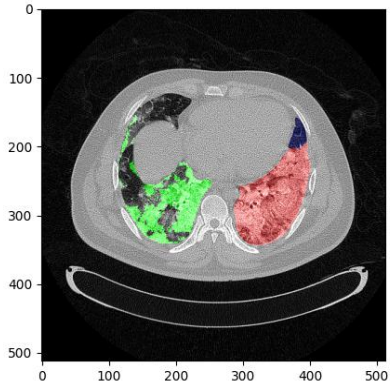
在构建完 FMED-Net 网络结构后，我们进行编程，实现了 FMED-Net 网络的搭建，并在 ARDS 患者的肺部 CT 影像数据集上划分并进行训练及测试，实现结果如表格：

Model	Dice Coefficient	Global correct	Average row correct			IoU			Mean IoU
U-Net	0.854	99.6	99.8	95.4	95.2	99.6	92.6	**93.6**	95.3
F^3Net	0.891	99.6	99.9	95.4	94.8	99.6	93.3	93.0	95.3
FMEDNet	**0.920**	99.6	**99.9**	**96.3**	95.2	99.6	**94.2**	93.2	**95.6**

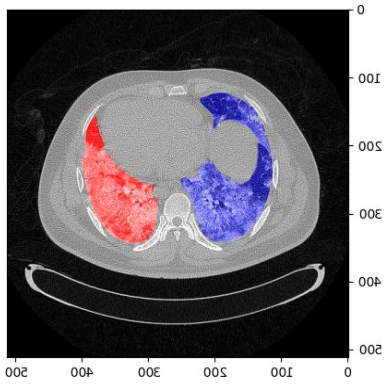
Table 1: Result of three model in a same test dataset

实验结果表明，fmednet 模型在各种指标上均有所提升，其中 DSC 提升近两个点。DSC 指标能很好地反映出模型分割出的边缘与实际分割边缘的重合程度，即模型对边缘细节分割的表现效果，DCS 指标的提升也能很好地体现 FMED-Net 模型的优势：由于精度损失较少，在细节处理上效果较好。

下图为模型效果演示，可以看到本模型能够很好地分割出图像的病变区域，相比传统的 U-net 有较大的精度提升。



图表 1 Our FMED-Net



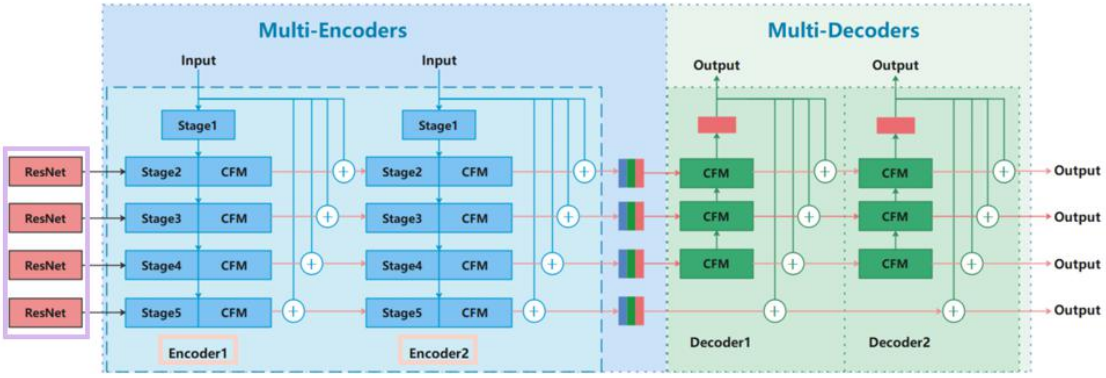
图表 2 Tranditional U-net

三、FMED-Net 的改进和延伸性实验

3.1 FMED-Net 的实际训练问题和 resnet 改进

在这一部分，我主要针对 FMED-Net 训练中的实际问题做了一些相应的改进：

我发现训练过程中最大的问题是模型参数量过大，且我们计算资源有限，这给我们带来了极大的困扰，经常出现 train 不动的情况；所以这里针对 resnet 模块，在原先每个 encoder 内部含有独立 Resnet 的基础上，我将其改进为了在所有 encoder 前设置一个统一的 resnet，以达到共享参数的效果。发现这样做，和之前同学的训练结果对比之下，模型效果不仅没有变差，反而大大加快了训练速度。（具体见后训练结果部分）



3.2 FMED-Net 的公共数据集训练结果

这里我采用了 RIDER Lung CT 数据集。RIDER Lung CT 数据集是用于评估非小细胞肺癌患者当天重复计算机断层扫描（CT）的肿瘤一维，二维和体积测量的变异性的数据集，共 15419 张图片。该研究纳入了 32 例非小细胞肺癌患者，每例患者在 15 分钟内通过相同的成像方案进行了两次胸部 CT 扫描。三位放射科医师在两次扫描中独立测量每个病变的两个最大直径，并且在另一个期间，在第一次扫描时测量相同的肿瘤。在单独的分析中，应用计算机软件来帮助计算两次扫描中每个病变的两个最大直径和体积。使用一致性相关系数(CCC)和 Bland-Altman 图来评估两次重复扫描的测量之间的协议(再现性）和相同扫描的两次重复读数之间的一致性（重复性）。该数据集于 2015 年由 TCIA(The Cancer Imaging Archive) 发布，具体介绍链接如下：

https://wiki.cancerimagingarchive.net/display/Public/RIDER+Lung+CT

我首先进行了进一步的数据清洗，解决了 image 和 mask 存在的部分不对应的情况。然后利用公共数据集数据量充分的优势，分别针对切片大小 1000、2000、3000 及以上，进行了较为细致的对比实验,在此分别截取此前 ARDS 患者的肺部 CT 影像数据集和公共数据集上 1000 和 3000 张影像具有代表性的实验结果数据如下：

```
[epoch: 4]
train_loss: 0.2398
lr: 0.009636
dice coefficient: 0.920
global correct: 99.6
average row correct: ['99.9', '96.3', '95.2']
IoU: ['99.6', '94.2', '93.2']
mean IoU: 95.6
```

```
[epoch: 14]
train_loss: 0.0570
lr: 0.000000
dice coefficient: 0.656
global correct: 94.4
average row correct: ['93.8', '98.2', '97.9']
IoU: ['93.6', '60.7', '78.3']
mean IoU: 77.5
```

```
[epoch: 26]
train_loss: 0.0980
lr: 0.000515
dice coefficient: 0.806
global correct: 99.3
average row correct: ['99.4', '98.8', '98.6']
IoU: ['99.2', '93.1', '96.6']
mean IoU: 96.3
```

经过对照实验，在此谨暂拟出以下结论：

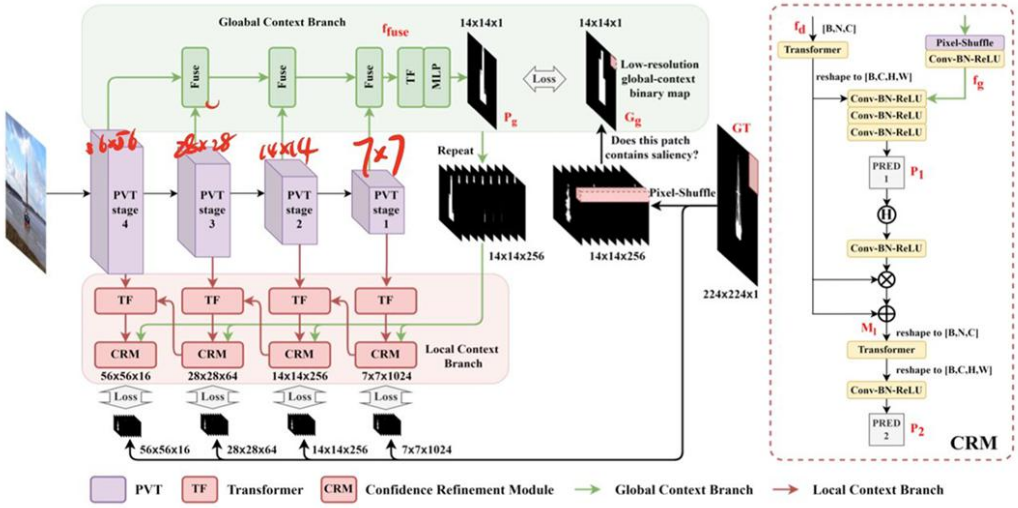
- 小数据集的收敛速度更快。ARDS 数据集和 1000 张影像的公共数据集在 15 个 epoch 左右就已经收敛，而 3000 张影像的切片在近 30 的 epoch 时 loss 还有明显的下降趋势。
- 数据的质量和精度在一定程度上来说比数据量更为重要。在不足一千张的 ARDS 数据集上我们得到了几乎最优的实验结果，特别是 dice coefficient 指标；而 mean IoU 方面 4 个 epoch 就可以和 3000 张影像 20 个 epoch 的结果相媲美，且训练速度是不可同日而语的。
- 在较大的数据集上更容易体现出任务的一些本质特点。例如在 1000 和 3000 张影像上，我明显地发现第二个 IoU 指标（即左肺分类）较低。目前我们的猜想是左肺部分临近心脏，器官较为复杂，所以任务难度较大；而这在大数据集上有明显的体现。
- 如上一条所说，在面对较难左肺分类任务时，大数据集的优势就可以比较明显的体现出来。即使在 20 个 epoch 以后，1000 张影像的左肺分类 IoU 结果依然停留在一个较低的指标上（70 左右），这远远低于 3000 张影像的分类结果。
- 在 dice 系数方面，很可能大数据集未必带来更好的结果。如上第一张和第三张图所示，ARDS 的 dice 系数比 3000 张影像的结果高了 0.1+，我的猜想是大数据集降低了预测正确结果的比重；换言之，大数据集上预测的精度反而下降了。

然而，由第三张图可以看出，总体上来说在大数据集上的训练结果还是能很好地凸显出我们模型的优势的，基本上达到了 state-of-art 的水平。

3.3 salient object detection 领域的延伸性探索和迁移

我受以上 F3N 的启发，还将该领域最新的 Sota 方法迁移到了医学图像领域，即前不久刚刚提出的 selfreformer 模型，模型榜单和结构图如下：

Salient Object Detection	DUT-OMRON	SelfReformer	max_F1	0.836	# 1
			MAE	0.041	# 1
			E-measure	0.886	# 1
			S-measure	0.856	# 1
Salient Object Detection	DUTS-TE	SelfReformer	MAE	0.026	# 1
			max_F1	0.916	# 1
			E-measure	0.920	# 1
			S-measure	0.911	# 4
Salient Object Detection	ECSSD	SelfReformer	MAE	0.027	# 1
			max_F1	0.957	# 1
			S-measure	0.935	# 1
			E-measure	0.928	# 1
Salient Object Detection	HKU-IS	SelfReformer	MAE	0.024	# 1
			E-measure	0.959	# 1
			max_F1	0.947	# 1
			S-measure	0.930	# 1



这里由于数据集数据类型不匹配的原因，我又清洗整理了 COVID-19 Radiography 数据集：

名称	标注内容	类型	模态	数量	标签格式	文件格式
covid19-radiography-database	新冠/其他肺炎/正常	分类	CT	219+1314+1345		图片

（具体介绍链接： COVID-19 Radiography Database | Kaggle ）
并将 self-reformer 迁移到此领域，在这个数据集上训练，也取得了优于此前分割方法的优越结果（稍逊于 FMED-Net）。这说明 Salient object detection 领域和医疗图像

领域有相通之处，未来有很多可以修改、融合的工作。

四、FMEDNet 的改进尝试以及局限性探索

4.1 F3Net 模型复现

在这一部分，我们先对 F3Net 的结构进行了代码层面的研究并且对论文进行了复现并在 ECSSD 数据集上测试了模型的效果。结果与论文基本一致。

	maxF	Fm	MAE	WFM	sm	em
F3Net	0.936	0.901	0.045	0.895	0.909	0.896

表 4-1 F3Net 在 ECSSD 数据集上的指标

部分结果图片和 Mask 比较如下：

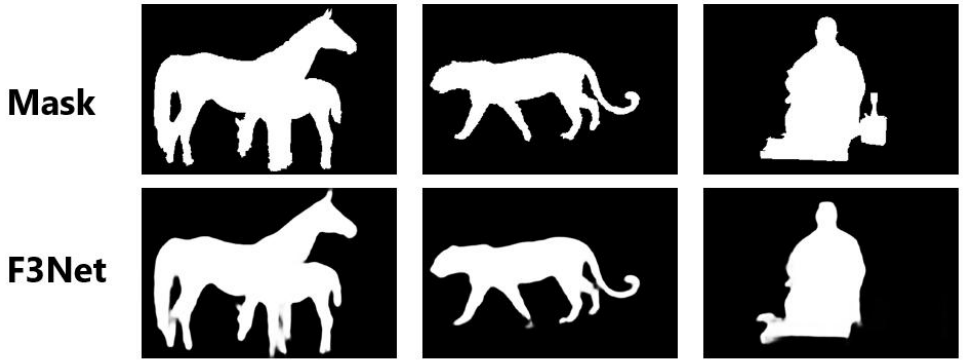


图 4-1 F3Net 结果图片与 Mask 的比较

4.2 提出 FMEDnet 模型变种——FMEDnet_Inv

在复现之后，我们感受到 F3Net 在融合特征方面的强大能力，对 F3Net 中的 CFM 和 CFD 模块有了更深的理解。然后，我们就对 FMEDnet 编码器部分精炼（图 4-2 红色框部分）的方向产生了思考。

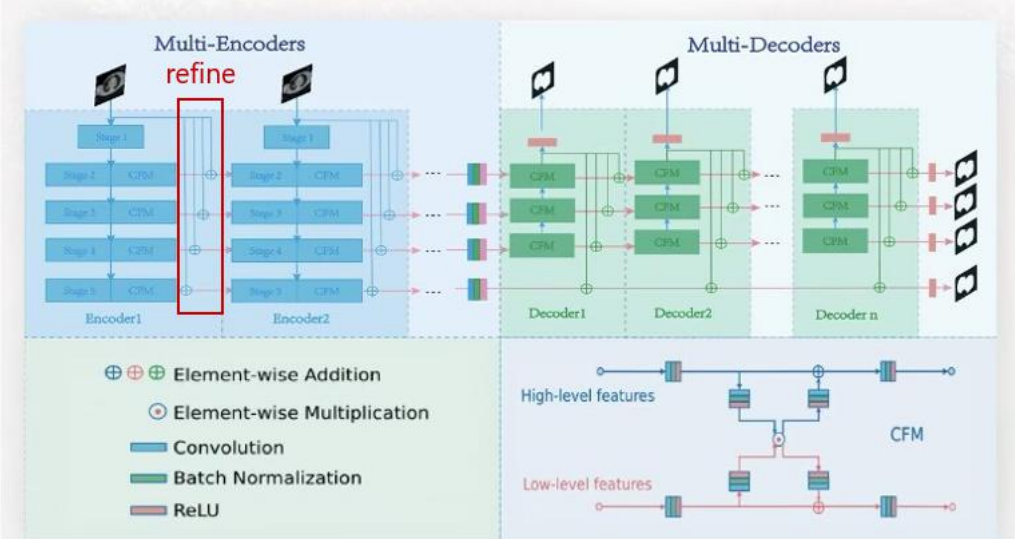


图 4-2 FMEDnet 编码器的精炼部分

编码器部分的精炼使用输入图片，即低级特征与高级别特征直接相加融合。这里的思想是由 F3Net 启发的。F3Net 的作者指出图像自上而下，语义特征逐渐增强，位置特征逐渐减弱。编码器部分使用输入图片进行精炼引导模型更多关注位置信息。但是，作为一个编码器，更多关注图片的语义信息是不是能取得更好的效果呢？于是，我们提出来一个变种。如图 4-3 所示，我们使用最高级别的特征与低层次特征相加的方式进行融合，引导模型更多关注图片的语义信息。我们将这一变种模型称为 FMEDnet_Inv，表示将自上而下的精炼方式反转过来改为自下而上的精炼方式。

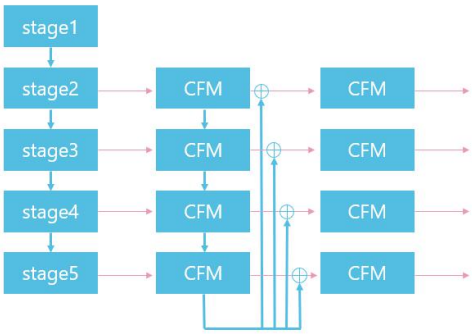


图 4-3 FMEDnet_Inv 的编码器部分

4.3 实验以及结果分析

我们在 ECSSD 数据集上测试比较了 F3Net、FMEDnet 以及 FMEDnet_Inv 三个模型。与之前的实验不同，ECSSD 并非医学影像的数据集，而是自然图片的数据集，其含有 1,000 张图片。实验中，我们使用 SGD 作为优化器，batch size 为 8，迭代的 epoch 数选为 32。实验中，我们将数据集按照 2:8 分成了测试集和训练集。实验在一张 2080Ti 上累积训练了 4 个小时。实验结果如下：

Model	maxF	Fm	MAE	WFM	sm	em
F3Net	0.936	0.901	0.045	0.895	0.909	0.896

FMEDnet	0.903	0.878	0.066	0.842	0.869	0.881
FMEDnet_Inv	0.383	0.340	0.362	0.283	0.405	0.562

表 4-2 F3Net，FMEDnet，FMEDnet_Inv 在 ECSSD 数据集上的指标。指标中，除了 MAE 指标越小说明效果越好外，其余指标均为越大效果越好

部分结果图片和 Mask 比较如下：

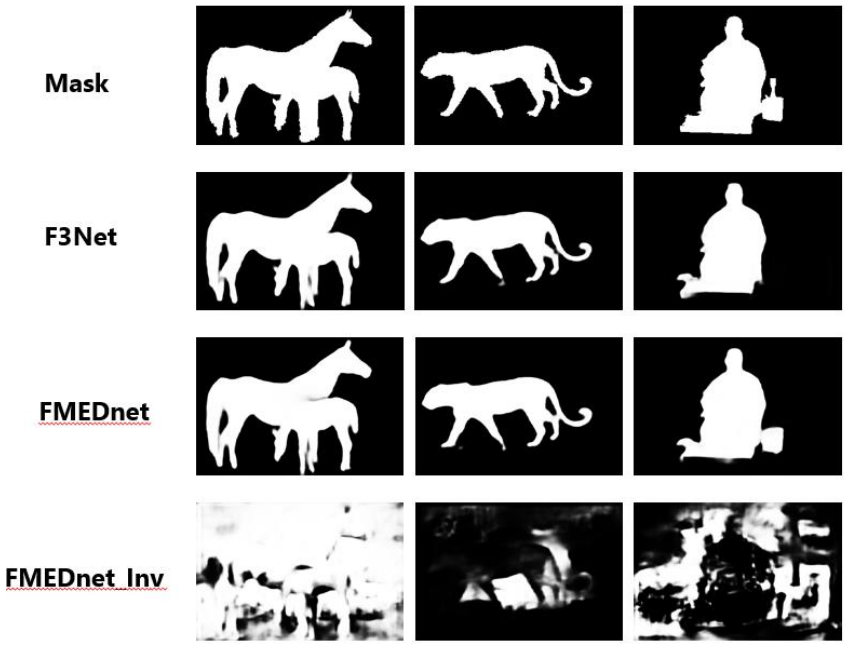


图 4-4 F3Net，FMEDnet，FMEDnet_Inv 结果图片与 Mask 的比较

从指标和测试结果可以看出 FMEDnet_Inv 模型思路是不正确的，是一次失败的尝试，我们分析这是因为高层次的语义信息极大干扰了物体的识别与训练。在 F3Net 和 FMEDnet 的比较中，我们发现，FMEDnet 的在各项指标上都比 F3Net 更低。这是不符合预期的。因为在肺部影像数据集上，FMEDnet 相较于 F3Net 时略胜一筹的。为了找出原因，我们比较了 FMEDnet 和 F3Net 分别在训练集和测试集的指标数据（表 4-3 和 4-4）：

F3Net	maxF	Fm	MAE	WFM	sm	em
Train	0.973	0.957	0.014	0.958	0.961	0.950
Test	0.936	0.901	0.045	0.895	0.909	0.896

表 4-3 F3Net 在训练集和测试集上的指标数据

FMEDnet	maxF	Fm	MAE	WFM	sm	em
Train	0.975	0.961	0.013	0.961	0.961	0.952
Test	0.903	0.878	0.066	0.842	0.869	0.881

表 4-4 FMEDnet 在训练集和测试集上的指标数据

比较 F3Net 和 FMEDnet 在训练集上的指标我们发现，FMEDnet 时超过 F3Net 的。由此我们得出结论：FMEDnet 相较于 F3Net 多出来的编码器模块使得 FMEDnet 陷入过拟合之中。由于医学影像数据集具有空间语义相对单一的特点，即在某一个样本的某一片区域表示左肺，那么其他样本的相同位置有很大可能仍表示左肺，所以 FMEDnet 中编码器部分导致的过拟合对于提升任务指标是有利的。反之，自然图像没有这样的特点，所以这样的过拟合导致了指标的下降。

五、总结

在第一部分，我们收集了肺部数据分割的公共数据集并进行预处理，以便后续工作进行。同时，引入 F3Net 模型，将之应用迁移到医学图像分割领域，与之后我们改进的模型进行进一步比较。

在第二部分，我们分析了 F3Net 的不足之处，改进了 F3Net，提出了多级编码器解码器对称的 FMED-Net 网络，并在病变严重的 ARDS 患者的肺部 CT 影像数据集上进行训练及测试，通过结果分析，我们发现 FMED-Net 网络在 F3N 的基础上，能进一步提高模型细节分割处的表现。

在第三部分，我们主要从训练效果层面改进了 FMED-Net，提出了共用 resnet 编码器参数的结构，并进行了较为详尽的对照实验，得出了一些阶段性结论；同时，还借鉴 selfreformer 模型，并将其迁移到医学图像分割领域，在 covid 相关数据集上取得了不错的效果。

在第四部分，我们对 F3Net 的结构进行了更加深入的理解，提出并测试了 FMEDnet 的不同变种的效果，探索了 FMEDnet 模型的局限性和特点。