arXiv:2303.00279v1 [eess.IV] 1 Mar 2023

# 基于视觉-语言对齐的 COVID-19 粗到细分割

*Dandan Shan*[⋆] *Zihan Li*[†] *WentaoChen*[‡] *Qingde Li*[*] *JieTian*[◇] *Qingqi Hong*[⋆††⋆]厦门大学 [†]伊利诺伊大学厄巴纳-香槟分校 [‡]北京邮电大学 [*]赫尔大学 [◇]中国科学院 [††]香港COCHE

**摘要**

新冠肺炎病灶分割有助于医生更好地诊断和治疗 COVID-19。然而，由于新冠肺炎数据集缺少详细信息和高质量标注，相关研究较少。为解决上述问题，我们提出了 C2FVL，这是一种基于视觉-语言对齐的粗到细分割框架，可融合包含病灶数量的文本信息与图像信息中的具体位置。引入文本信息使网络能够在具有挑战性的数据集上取得更好的预测结果。我们在两个 COVID-19 数据集（包括胸部X光和 CT 数据集）上进行了大量实验，结果表明，我们提出的方法优于其他最先进的分割方法。

索引词— 由粗到细、视觉语言、语义分割

## 1. 引言

深度学习技术已广泛应用于医学图像处理领域，利用神经网络进行病灶分割可以显著减少时间和劳动力成本。由于多模态图像比单模态图像包含更多相关信息 [1]，一些研究 [2, 3]将不同模态的图像融合起来，以分割目标。然而，医学图像缺乏详细信息，仅凭医学图像进行精确分割具有挑战性。一些工作 [4, 5, 6] 采用图像信息与文本信息融合的方法进行图像分割。尽管Tomar等人 [7] 将图像信息和文本信息融合应用于医学图像分割，但未考虑目标的位置信息。病灶区域的分割是诊断和治疗 COVID-19的有效手段。然而，只有经验丰富的放射科医生才能准确标注病灶，而公开可用的数据集较少 [8], , 这使得分割网络的精度难以提高[9, 10, 11]。为解决上述问题，我们在COVID-19分割过程中引入文本信息。由于文本包含病灶的数量和具体位置信息，因此可以帮助网络在有限且标注不完善的数据集上，从粗到细地学习具有丰富语义信息的特征，从而学习具有丰富语义信息的特征。总而言之，本文的主要贡献如下:

- 我们采用卷积神经网络和视觉 Transformer 构建了 C2FVL 分割框架，以有效实现对 COVID-19 的精准分割。
- 我们提出了视觉语言对齐模块（VLAB）和一种新颖的损失函数，以促进文本与图像信息的对齐，从而提高分割精度。
- 我们在两个 COVID-19 数据集上，将 C2FVL 与最先进的分割方法进行了比较。实验结果表明，C2FVL 的性能最佳。我们提出的 C2FVL 代码已发布于 GitHub[1]。

## 2. 相关工作

**多模态信息融合：** 多种模态能够提供互补信息，将它们融合可以有效提高医学图像分割的准确性。Dolz 等人 [2] 利用不同模态之间的密集连接交换信息。CLIP [4] 采用对比学习方法学习图像和文本信息，实现了零样本迁移学习。TGANet [7] 引入了关于息肉大小和数量的文本信息，从而能够自动分割不同尺寸的息肉。

**COVID-19分割：** Saeedizadeh 等人 [10] 在 UNet 的损失函数中加入正则化项，以检测 CT图像中感染 COVID-19 的胸部区域。Alom 等人 [11] 在其设计的系统中使用 NABLA-N 模型分割 COVID-19感染区域。

## 3. 方法

### 3.1. C2FVL模型结构

如图 1 所示，我们的模型采用编码器-解码器结构，其中编码器使用分层卷积神经网络和视觉 Transformer 提取粗略的图像特征。在跳跃连接处，将文本信息与图像信息融合，并设计 VLAB 模块，以促进文本与图像之间的跨模态对齐。解码器由

前两位作者贡献相同。通讯作者：Qingqi Hong。

[1]https://github.com/HUANGLIZI/C2FVL

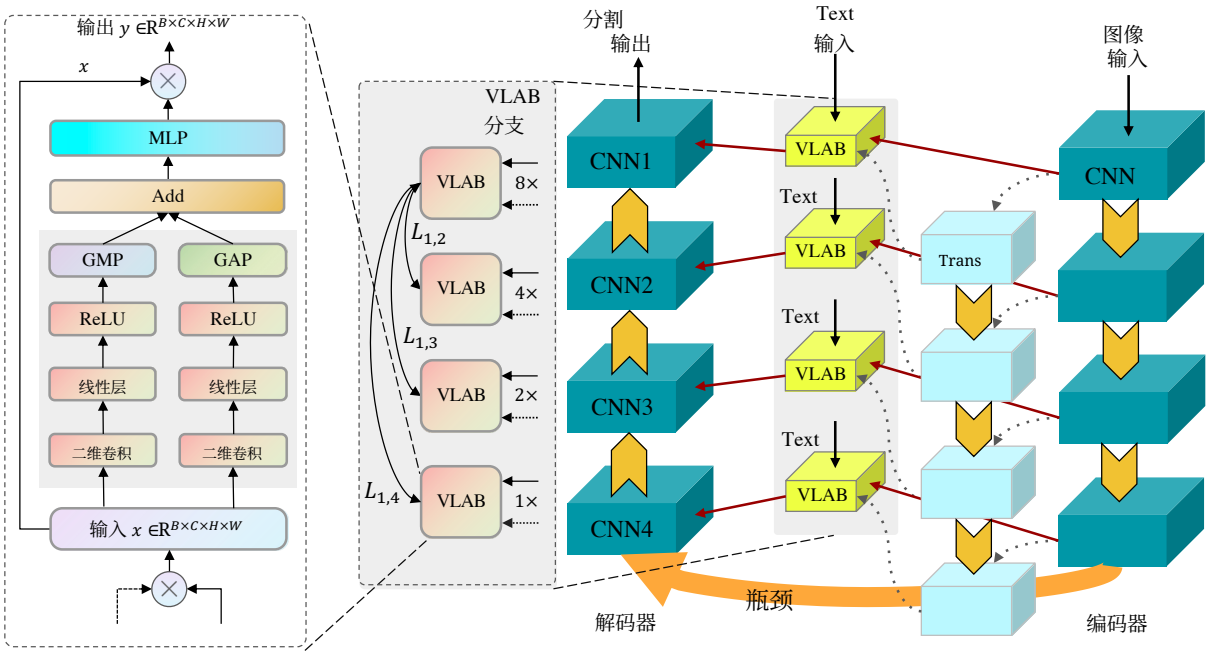


**图1**。C2FVL框架概览。它由三个部分组成：编码器、VLAB分支和解码器。

用于处理低层细节信息的卷积神经网络组成，并将对齐后的文本-图像特征用于重建精细的分割掩码。

### 3.2. 编码器模块

对于医学图像分割任务，全局特征和局部特征都至关重要。传统卷积神经网络使用固定大小的感受野来学习图像的局部特征。 [12] 使用平均池化提取全局特征，但其能力有限。相比之下，视觉 Transformer 中的多头自注意力机制直接作用于整幅图像，因此能够学习更多全局特征。因此，我们在编码器模块中使用若干卷积神经网络块和视觉 Transformer 块，共同提取图像特征。首先，将图像输入卷积神经网络块，然后经过卷积操作、批归一化和激活函数处理后，将特征图送入视觉 Transformer。同时，使用最大池化对特征图进行下采样，以保留更多纹理特征。对于激活函数，我们使用 ReLU 来增强网络的表达能力。在视觉 Transformer 部分，对卷积神经网络提取的特征进行编码，以捕获全局特征。在每个阶段，调整视觉 Transformer 输出的尺度，使其与卷积神经网络块的输出尺度相同，并将二者相加，以融合局部特征和全局特征，如下列公式所示。

$$F_{rt,i} = \sigma\left(BatchNorm\left(Conv\left(Upsample\left(F_{vit,i}\right)\right)\right)\right) \quad (1)$$

$$F_{encoder,i} = F_{cnn,i} + F_{rt,i} \quad (2)$$

其中，$F_{vit,\ i}$ 表示视觉 Transformer 第 $i$ 层的输出，$F_{rt,\ i}$ 表示$F_{vit,\ i}$的重建结果，而$F_{cnn,\ i}$ 表示 CNN 第 $i$ 层的输出， $\sigma$ 是 ReLU激活函数。

### 3.3. 多尺度视觉-语言聚合

我们将文本信息转换为维度为 8 的向量。第一维表示病灶区域是否为双侧，第二维表示病灶数量，之后每三维依次表示病灶在左肺和右肺中的位置。例如，假设文本信息为“肺部双侧感染，两个感染区域，左肺中上部和右肺中下部”。在这种情况下，我们将其转换为向量 $V_{text} = [1, 2, 1, 1, 0, 0, 1, 1]$。不同于 [7], 使用文本引导的图像分类来辅助分割任务，我们使用文本信息来选择特征。具体而言，对于不同尺度的特征图，我们采用重复操作，使其与图像特征的通道数对齐，然后利用文本对图像特征进行加权，以提取病灶的位置和数量特征，同时抑制其他无关特征，如公式（3）所示。

$$F_{vl,i} = F_{encoder,i} \odot \left(V_{text}.repeat\left(channel_i/8, 1\right)\right) \quad (3)$$

### 3.4. 视觉-语言对齐模块（VLAB）

受卷积块注意力模块 [13] 和 CUM [14], 的启发，我们设计了具有并行结构的视觉语言对齐模块（VLAB），以促进文本特征与图像特征的对齐。如图 1 所示，在左、右两个分支中，首先使用多层感知机结构处理输入特征，然后分别对特征图进行全局最大池化（GMP）和全局平均池化

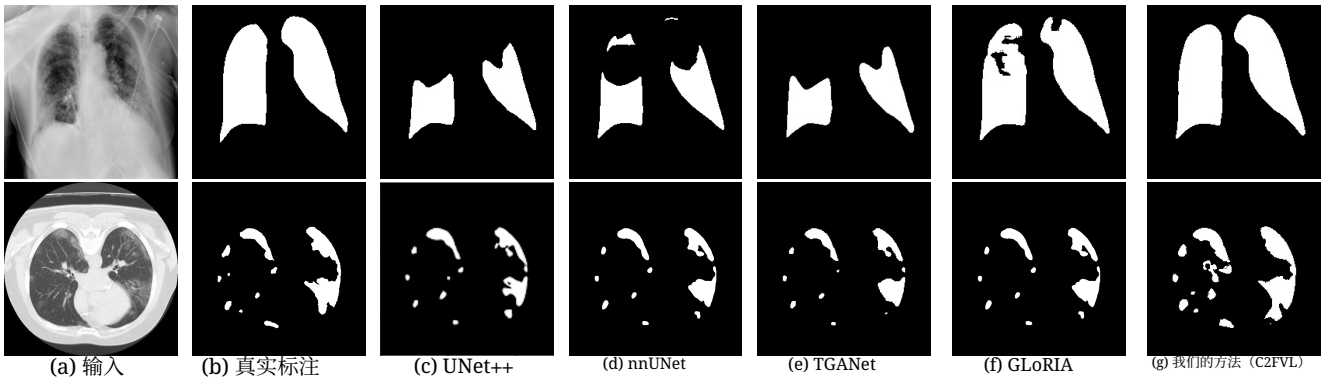


**图2**。QaTa-COVID 数据集（第1行）和 MosMedData+ 数据集（第2行）上的可视化结果。从左到右依次为：(a) 输入图像，(b) 真实标注，(c) UNet++ 和 (d) nnUNet，它们是未使用文本信息的基线方法的预测结果；而 (e) TGANet 和 (f) GLoRIA 是使用文本信息的预测结果。(g) 是我们提出的 C2FVL 的预测结果。

（GAP）。将两个分支的输出相加，以进一步增强相同的特征，最后与经过两层多层感知机结构处理后的输入相乘。该过程可定义如下。

$$F_{avg} = GlobalAvgPool\left(MLP\left(x\right)\right) \tag{4}$$

$$F_{max} = GlobalMaxPool\left(MLP\left(x\right)\right) \tag{5}$$

$$y = MLP\left(F_{avg} + F_{max}\right) \otimes x \tag{6}$$

此外，我们还计算不同 VLAB输出之间的余弦损失，以促使每个网络层关注相同的病灶区域，并将这部分损失添加到 Dice损失 $L_{Dice}$ 和交叉熵损失 $L_{CE}$之后。所设计的损失函数如公式 (7) 和 (8) 所示。

$$L_{4,i} = 1 - \frac{y_4 \cdot Downsample(y_i)}{|y_4| \times |Downsample(y_i)|} \tag{7}$$

$$L_{V1} = \frac{1}{2}L_{Dice} + \frac{1}{2}L_{CE} + \alpha L_{4,1} + \beta L_{4,2} + \gamma L_{4,3} \tag{8}$$

其中，$L_{4,\ i}(i \in 1,\ 2,\ 3)$ 表示第 4 层 VLAB输出 $y_4$ 与第 $i$ 层 VLAB输出 $y_i$之间的余弦损失。$\alpha$、$\beta$、$\gamma$ 的默认值均为 0.5。此外，我们还设计了 $L_{1,\ i}(i \in 2,\ 3,\ 4)$，用于计算第一层 VLAB 的输出与较低层 VLAB 输出之间的余弦损失，如公式 (9) 和 (10) 所示。

$$L_{1,i} = 1 - \frac{y_1 \cdot Upsample(y_i)}{|y_1| \times |Upsample(y_i)|} \tag{9}$$

$$L_{V2} = \frac{1}{2}L_{Dice} + \frac{1}{2}L_{CE} + \alpha L_{1,4} + \beta L_{1,3} + \gamma L_{1,2} \tag{10}$$

## 4. 实验

### 4.1. 设置

**数据集：** 我们在实验中使用开放获取的 QaTa-COVID 数据集 [15] 和 MosMedData+ 数据集 [16, 17]来评估性能。QaTa-COVID 的训练集和测试集分别包含 7145 张和 2113 张带有标注的胸部X光图像。我们按 4:1 的比例将初始训练集划分为训练集和验证集。MosMedData+ 包含 2729 张肺部感染的 CT扫描切片。为平衡病灶的不同形态，我们将具有四种种前缀的图像视为四个类别。对于每个类别，我们按照 8:1:1 的比例将图像分配到训练集、验证集和测试集中。

**实现细节：** 我们的实验在四块 NVIDIA GeForce GTX 1080 Ti GPU 上进行，模型训练的初始学习率设为 1e-3，批大小设为 4。总迭代轮数设为 2000，早停轮数设为 50。

**表1.** 我们的方法（C2FVL）与其他最先进方法在数据集上的性能比较。

| 方法 | Text | QaTa-COVID | | MosMedData+ | |
|---|---|---|---|---|---|
| | | Dice（%） | 交并比（%） | Dice（%） | 交并比（%） |
| UNet [18] | × | 79.02 | 69.46 | 64.60 | 50.73 |
| UNet++ [19] | × | 79.62 | 70.25 | 71.75 | 58.39 |
| AttUNet [20] | × | 79.31 | 70.04 | 66.34 | 52.82 |
| nnUNet [21] | × | 79.89 | 70.58 | 72.59 | 60.36 |
| TransUNet [22] | × | 78.63 | 69.13 | 71.24 | 58.44 |
| Swin-UNet [23] | × | 77.27 | 67.96 | 63.29 | 50.19 |
| UCTransNet [24] | × | 79.15 | 69.60 | 65.90 | 52.69 |
| TGANet [7] | ✓ | 79.87 | 70.75 | 71.81 | 59.28 |
| GLoRIA[12] | ✓ | 79.94 | 70.68 | 72.42 | 60.18 |
| **Ours** | ✓ | **83.40** | **74.62** | **74.56** | **61.15** |

### 4.2. 与最先进方法的比较

我们在两个数据集上与最先进方法进行了对比实验，包括仅使用图像输入的方法和使用图像-文本输入的方法。如图2所示，第一行表明，其他方法在肺部上方区域出现了不同大小的缺失，而我们提出的 C2FVL 模型能够较为完整地分割病灶。如第二行所示，对于对诊断更为重要的右肺区域，C2FVL 预测的结果更接近真实标注。如表1所示，与表现最佳且不使用文本的 nnUNet 方法相比，C2FVL 在 QaTa-COVID 数据集上的 Dice 分数和交并比分别提高了 3.51% 和 4.04%。同时，在 MosMed-Data+ 数据集上，Dice 分数和交并比分别提高了 1.97% 和 0.79%。与采用文本的基线模型（如 GLoRIA） [12], 相比，C2FVL 在这两个数据集的两项指标上均有所提升，提升幅度分别为 3.46%、3.94%、2.14% 和 0.97%。

### 4.3. 消融研究

我们在 QaTa-COVID 数据集上，针对相关超参数和模型的不同组件进行了一系列消融实验。

**不同超参数的消融研究：**对于批大小，我们选取 2、4 和 8 三个值进行比较；对于学习率，我们设置为 1e-4、1e-3 和 1e-2。从表 2 的结果可以看出，当批大小为 4、学习率为 1e-3 时，模型的 Dice 分数和 IoU 值最佳。对于损失系数，我们从 $\alpha = 0.5$、$\beta = 0.5$、$\gamma = 0.5$ 开始，在 $\alpha$、$\beta$、$\gamma$ 中固定两个值为 0.5，并改变另一个值，以观察模型性能的变化趋势。如图 3 所示，最优参数组合为 $\alpha = 0.5$、$\beta = 0.5$、$\gamma = 0.5$。

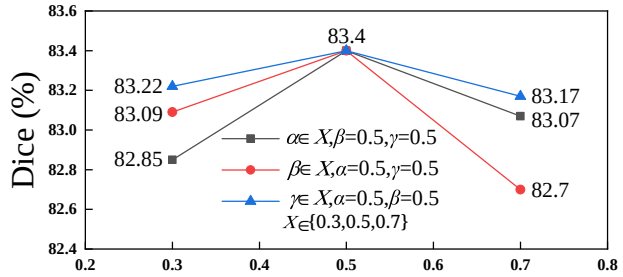


**图 3**。不同余弦损失系数下的消融研究。灰色、红色和蓝色线条分别表示以 $\alpha$、$\beta$ 和 $\gamma$ 为变量时 Dice 的变化趋势。

**表 2**。不同超参数下的消融研究。

| 超参数 | 值 | QaTa-COVID Dice (%) | QaTa-COVID 交并比（%） |
|---|---|---|---|
| 批大小 | 2 | 82.03 | 73.32 |
| | 4 | **83.40** | **74.62** |
| | 8 | 83.32 | 74.57 |
| 学习率 | 1e-4 | 82.93 | 74.19 |
| | 1e-3 | **83.40** | **74.62** |
| | 1e-2 | 82.83 | 73.92 |

**表3**。在 QaTa-COVID 数据集上对不同组件有效性的消融研究。

| 方法 | CNN | 文本输入 | | C2FV损失 | | VLAB | Dice（%） | 交并比（%） |
|---|---|---|---|---|---|---|---|---|
| | | 单一 | 多重 | $L_{V1}$ | $L_{V2}$ | | | |
| UNet++ | ✓ | | | | | | 79.62 | 70.25 |
| C2FVL | ✓ | ✓ | | | | | 83.01 | 74.26 |
| | ✓ | | ✓ | | | | 83.18 | 74.43 |
| | ✓ | | ✓ | ✓ | | ✓ | 83.08 | 74.42 |
| | ✓ | | ✓ | | ✓ | ✓ | **83.40** | **74.62** |

**不同组件有效性的消融研究：**如表3所示，我们使用提取图像特征的卷积神经网络作为骨干网络，然后逐步加入文本信息和 VLAB。值得注意的是，与骨干网络相比，在原始图像输入中加入单一文本后，模型性能显著提升，Dice 分数提高了 3.39%，交并比提高了 4.01%。在每一层加入相应的尺度文本后，模型性能进一步提升。在此基础上，采用下采样方法的 VLAB 余弦损失性能有所下降，但采用上采样方法的 VLAB 余弦损失，其 Dice 分数和交并比均达到最优。因此，在后续实验中，我们对 VLAB 使用 $L_{V2}$。

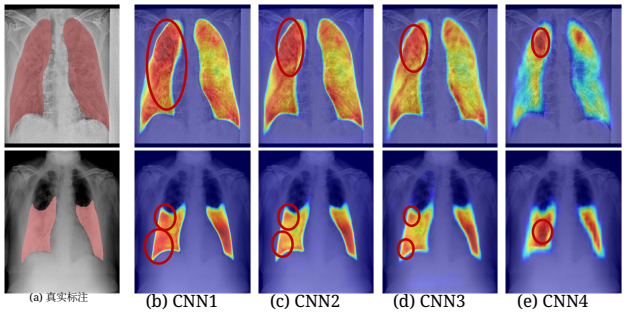


**图4**。QaTa-COVID 数据集可解释性研究的显著性图。从 (a) 到 (d) 分别是四个卷积神经网络解码器各自不同输出对应的显著性图，顺序为从底部（CNN4）到顶部（CNN1）。（e）“GT”表示对应的真实标注。

### 4.4. 可解释性研究

如图4所示，为验证我们的余弦损失能够增强网络对病灶区域的关注，我们使用 Grad-CAM [25]展示了 QaTa-COVID 数据集上解码器部分的激活区域。从第一行的左肺可以看出，CNN4 的激活映射显示出一片相对狭窄的强激活区域。随着网络层逐渐变浅，强激活区域开始从 CNN4 所在位置向邻近区域扩散，最终聚焦于整个病灶区域。对于第二行的左肺，CNN4 的激活图显示强激活区域主要位于病灶区域中心，随后扩展至左肺边缘，再从边缘向中心区域扩散。这一结果也表明，C2FVL 中设计的余弦损失能够帮助模型检测关注位置不准确的情况，并促使模型重新聚焦于正确的病灶区域。

## 5. 结论

本文提出了一种名为 C2FVL 的新型分割网络，该网络在训练过程中利用文本和图像信息。此外，我们设计了 VLAB，以促进图像信息与文本信息的对齐，并在损失函数中加入不同层之间的余弦损失，使网络的每一层都关注病灶区域。实验结果表明，与其他先进模型相比，我们的模型性能有所提升。尽管目前的工作只是初步探索了用于图像分类的多尺度视觉 Transformer，但我们预计未来将有更多研究开发 C2FVL，以应用于更多场景，包括更多新发布的数据集。

## 6. 致谢

本研究部分获得中国福建省自然科学基金（编号：2020J01006）、北京航空航天大学虚拟现实技术与系统国家重点实验室开放项目（编号：VRLAB2022AC04）以及 InnoHK 的资助。

## 7. 参考文献

[1] Binny Naik 等人，“机器学习与阿尔茨海默病多模态诊断分类的结论”，工业、生物医学与艺术视觉计算，第3卷，第1期，第1–18页，2020年。[2] Jose Dolz、Karthik Gopinath 等人，“Hyperdense-net：一种用于多模态图像分割的超密集连接卷积神经网络”，*IEEE*医学影像汇刊，第38卷，第5期，第1116–1126页，2018年。[3] Xiaohang Fu、Lei Bi 等人，“用于多模态PET-CT肺肿瘤分割的多模态空间注意力模块”，*IEEE*生物医学与健康信息学杂志，第25卷，第9期，第3507–3516页，2021年。[4] Alec Radford、Jong Wook Kim 等人，“通过自然语言监督学习可迁移的视觉模型”，载于国际机器学习会议。机器学习研究论文集，2021年，第8748–8763页。[5] Jiasen Lu、Dhruv Batra 等人，“ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表示”，神经信息处理系统进展，第32卷，2019年。[6] Zihan Li 等人，“LVIT：语言与视觉变换器在医学图像分割中的结合”，*arXiv*预印本 *arXiv:2206.14718*，2022年。[7] Nikhil Kumar Tomar、Debesh Jha、Ulas Bagci 等人，“TGANet：用于改进息肉分割的文本引导注意力”，*arXiv*预印本 *arXiv:2205.04280*，2022年。[8] Sneha Kugunavar 等人，“用于冠状病毒病疫情诊断与预后的卷积神经网络”，工业、生物医学与艺术视觉计算，第4卷，第1期，第1–14页，2021年。[9] Fei Xu 等人，“MRDFF：一种基于深度森林的CT全心脏分割框架”，方法，第208卷，第48–58页，2022年。

[10] Narges Saeedizadeh 等人，“COVID TV-UNet：使用施加连通性的 UNet 分割 COVID-19 胸部CT图像”，生物医学计算方法与程序更新，第 1 卷，第 100007 页，2021年。

[11] Md Zahangir Alom、MM Rahman、Mst Shamima Nas-rin、Tarek M Taha 和 Vijayan K Asari，“Covid mtnet：采用多任务深度学习方法进行新冠肺炎检测”，*arXiv* 预印本 *arXiv:2004.03747*，2020年。

[12] Shih-Cheng Huang、Liyue Shen 等，“GLoRIA：一种面向标签高效医学图像识别的多模态全局-局部表示学习框架”，载于国际计算机视觉大会论文集，2021年，第3942–3951页。[13] Zihan Li 等，“TFCNS：一种用于医学图像分割的卷积神经网络-Transformer混合网络”，载于人工神经网络国际会议 *2022*。施普林格，2022年，第781–792页。[14] Zihan Li 等，“Semi-WTC：一种通过权重-任务一致性进行攻击分类的实用半监督框架”，*arXiv* 预印本 *arXiv:2205.09669*，2022年。[15] Mehmet Yamac 等，“基于卷积稀疏支持估计器的X射线图像新冠肺炎识别”，*IEEE*神经网络与学习系统汇刊，第32卷，第5期，第1810–1820页，2021年。[16] Sergey P Morozov、AE Andreychenko 等，“Mosmed-data：含有新冠肺炎相关发现的胸部CT扫描数据集”，*arXiv* 预印本 *arXiv:2005.06465*，2020年。[17] “新冠肺炎CT分割数据集”，[EB/OL], http://medicalsegmentation.com/covid19/ 访问日期：2020年12月23日。[18] Olaf Ronneberger 等，“U-Net：用于生物医学图像分割的卷积网络”，载于国际医学图像计算与计算机辅助手术大会 *2015*。施普林格，2015年，第234–241页。[19] Zongwei Zhou 等，“UNet++：重新设计跳跃连接，以在图像分割中利用多尺度特征”，*IEEE*医学影像汇刊，第39卷，第6期，第1856–1867页，2019年。[20] Ozan Oktay、Jo Schlemper 等，“注意力U-Net：学习关注胰腺所在位置”，*arXiv* 预印本 *arXiv:1804.03999*，2018年。[21] Fabian Isensee、Paul F Jaeger 等，“nnU-Net：一种用于基于深度学习的生物医学图像分割的自配置方法”，《自然·方法》，第18卷，第2期，第203–211页，2021年。[22] Jieneng Chen、Yongyi Lu、Qihang Yu 等，“TransUNet：Transformer为医学图像分割提供强大的编码器”，*arXiv* 预印本 *arXiv:2102.04306*，2021年。[23] Hu Cao、Yueyue Wang、Joy Chen 等，“Swin-UNet：类U-Net的纯Transformer医学图像分割网络”，*arXiv* 预印本 *arXiv:2105.05537*，2021年。[24] Haonan Wang、Peng Cao 等，“UCTransNet：从通道级视角借助Transformer重新思考U-Net中的跳跃连接”，载于人工智能促进协会年会论文集，2022年，第36卷，第2441–2449页。[25] Ramprasaath R Selvaraju 等，“Grad-CAM：通过基于梯度的定位对深度网络进行可视化解释”，载于国际计算机视觉大会论文集，2017年，第618–626页。