arXiv:2301.02228v3 [eess.IV] 3 Apr 2023

MedKLIP：放射学中的医学知识增强语言-图像预训练

Chaoyi Wu[1,2], Xiaoman Zhang[1,2], YaZhang[1,2], Yanfeng Wang[1,2], Weidi Xie[1,2]l合作式Medianet创新中心，上海交通大学 [2]上海人工智能实验室{wtzxxxwcy02, xm99sjtu, ya zhang, wangyanfeng, weidi}@sjtu.edū.cn https://chaoyi-wu.github.io/MedKLIP/

摘要

本文探讨如何利用日常放射学实践中的配对图像-文本报告，通过领域专属知识增强医学视觉语言预训练（*VLP*）。具体而言，我们的贡献如下：**首先**，不同于直接处理原始报告的现有工作，我们采用一种新颖的三元组提取模块来提取医学相关信息，避免语言语法带来的不必要复杂性，并增强监督信号；**其次**，我们提出一种新颖的三元组编码模块，通过查询知识库进行实体翻译，以利用医学领域丰富的领域知识，并在语言嵌入空间中隐式构建医学实体之间的关系；**第**三，我们提出采用基于*Transformer*的融合模型，在图像块级别将实体描述与视觉信号进行空间对齐，从而实现医学诊断能力；**第四**，我们进行了全面实验以验证该架构的有效性，并在多个公开基准例如 *ChestX-ray14*、*RSNA* 肺炎、*SIIM-ACR* 气胸、*COVIDx CXR-2*、*COVID Rural* 和 *EdemaSeverity* 上进行基准测试。在零样本和微调设置下，与以往方法相比，我们的模型在疾病分类和定位方面均表现出优异性能。

1. 引言

随着深度学习的快速发展，许多研究相继提出，旨在推动医学领域的计算机辅助诊断 [44, 20, 53, 19]。尽管这些模型取得了巨大进展，但通常只能识别或分割某个封闭的解剖结构或疾病类别集合中的结构。一旦出现新的关注疾病，就需要耗费大量资源进行数据标注和模型重新训练，这从根本上限制了模型的实际应用价值。作为替代方案，近期

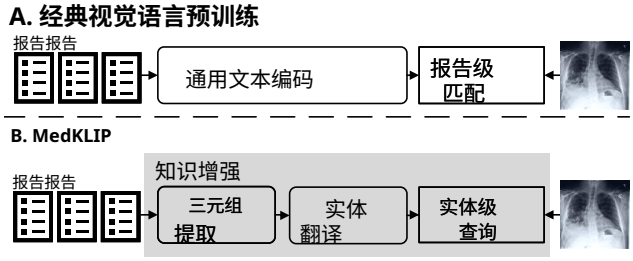


图1：我们的方法主要考虑将医学知识与视觉语言预训练相结合。我们提出了三元组提取和实体翻译模块，使网络能够通过细致的实体级信号进行监督。

研究考虑使用日常临床流程中生成的大量多模态数据构成的语料库来训练模型。例如，最常见的案例是包含X射线图像及其配对放射学报告的数据集 [18, 28, 31]。

本文介绍了我们在医学领域视觉语言表征学习方面的初步研究，旨在更好地实现零样本疾病诊断（分类）和定位。毫无疑问，计算机视觉领域也对这些任务进行了广泛研究，过去几年在开发基础模型方面取得了显著进展，例如 CLIP [48], ALBEF [33], BLIP [32]，等。然而，要在医学领域实现这一目标，必须解决不同的挑战，这需要学界投入研究力量：首先，数据可用性。在计算机视觉领域，训练基础模型通常需要数百万个图像-文本对，而医学领域只有几十万个可用样本 [31]。数据有限，使语言模型难以理解自由形式的报告 [6]。其次，计算机辅助诊断所关注的问题天然具有细粒度特征，需要辨别医学概念以理解疾病，因此领域知识至关重要；第三，鲁棒性至关重要，因此最好

具备可解释性，使诊断结果同时提供视觉定位，帮助放射科医生理解系统，并建立人类与机器之间的信任。

现有医学视觉语言预训练（Vision-Language Pre-training） [66, 45, 25, 6] 采用一种直接的训练范式，将原始报告与图像扫描进行匹配，如图1A所示，却忽略了医学先验知识。因此，我们提出了一种新颖的知识增强视觉语言模型，如图1B所示。首先，我们提出三元组提取模块，从原始报告中提取有用的医学实体（关键词），并将每份报告简化为一组三元组，表示为 {实体，位置，存在}。将报告分解为三元组，能够利用报告中的结构先验，以尽可能少的信息损失有效地表示报告；其次，我们借助定义完善的医学词汇知识库，将医学实体转换为细粒度描述，该知识库倾向于用常见词汇解释疾病。因此，计算这些描述的文本嵌入能够隐式地建立医学实体之间的关系；第三，我们将实体视为查询集，并采用基于Transformer的架构，使图像块与实体描述对齐，从而在实体级别提供显式监督信号。因此，我们能够借助空间热力图形式的视觉证据，同时推断特定疾病的可能性，即提供粗略定位以增强可解释性。

我们在一个广泛使用的医学图像-报告数据集 MIMIC-CXR [31]，上对模型进行预训练，并在众多公开基准上严格评估其疾病诊断能力，例如例如，ChestX-ray14 [56], RSNA 肺炎 [49], SIIM-ACR 气胸 [1], COVIDx CXR-2 [46], COVID Rural [52, 15], 和 EdemaSeverity [7]。在涵盖不同图像分布的多种疾病上，我们的模型在零样本分类和定位任务中取得了最先进的性能；经过进一步微调后，模型仍显著优于以往模型。

2. 相关工作

通用视觉-语言预训练（VLP）模型。在计算机视觉领域，此类研究路线近年来取得了巨大成功。总体而言，已开发的架构要么采用双流方法，也就是双编码器 [4, 33, 30],即，要么基于单流方法 [35, 9], ，后者侧重于视觉-语言融合。特别是，有几项研究 [13, 37, 64] 考虑将常识知识融入视觉-语言预训练；然而，本文聚焦于医学领域，该领域显然更加精细，也需要更多专业知识。

医学命名实体识别（NER）模型。各种自然语言处理（NLP）方法已经被提出用于从放射学报告中提取信息 [47, 28, 42, 50]。这些早期方法只考虑疾病，因此造成了信息损失。后续最先进的研究 [29, 59]提出了无需预先定义封闭疾病集即可提取不同实体之间关系的方法，从而以较高准确率保留了大部分有用信息。在弱监督 [65] 和报告生成领域 [14], , NER 方法已展现出重大影响，并极大地启发了我们：通过注入医学领域知识，实现更有效的视觉-语言预训练。

医学知识增强模型。利用外部医学知识增强深度学习模型并非新课题 [60]。根据医学知识的使用方式，这些方法可分为基于模型的方法和基于输入的方法。在基于模型的方法中，作者旨在模仿放射学或诊断实践来设计模型 [34, 23, 55, 26, 43, 21, 12]。而在基于输入的方法中，知识被视为额外输入，用于特征计算 [63, 61, 51, 11]，或用于指导最终的训练损失 [8, 27, 36, 38, 41], ，这通常应用于报告生成任务 [57, 62, 39, 11]。然而，这些工作都没有针对医学领域中基于图像-报告的视觉-语言预训练。

医学视觉语言预训练领域的同期工作。现有的医学视觉语言预训练方法沿用双流架构 [66, 25, 7, 45, 58,10], 即采用对比学习，且不包含融合模块。例如，ConVIRT [66] 最初提出将对比损失作为代理任务，以对齐医学扫描图像及其对应报告；随后，LoVT 和 GLoRIA 专注于提升局部对齐性能 [25, 7]。BioViL 注意到报告中的语言模式不同于自然文本，因此重新设计了语言模型 [6]。Align [10]提升了视觉语言预训练模型在医学视觉问答（VQA）任务上的性能。近期的 arXiv 预印本 MedCLIP [58], 则考虑利用未配对数据来弥补数据稀缺。与我们最相关的是 CheXzero [54], , 其目标是实现零样本诊断。尽管现有工作已取得重要进展 [66, 25, 7, 45, 58, 54], ，但它们仍将医学文本和图像视为普通自然数据，并未明确利用医学领域丰富的先验知识。本文考虑融入领域知识，精心重新设计预训练流程，并致力于实现对X射线扫描的准确诊断。

3. 方法

本节首先在第3.1节介绍所考虑的问题场景，接着在第3.2节说明我们的报告预处理操作和三元组提取。然后，我们在第3.3节介绍所提出的知识增强架构，包括视觉编码器、知识增强三元组编码器以及用于对齐

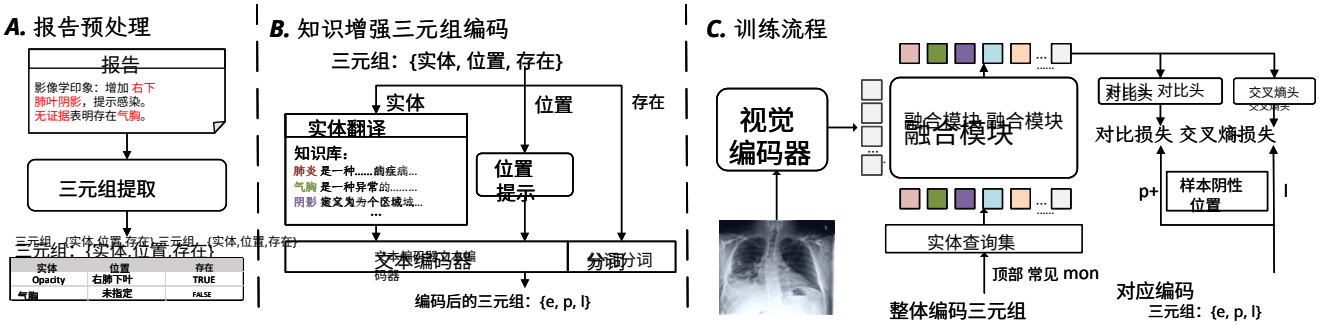


图2：我们方法的整体框架。首先，我们利用三元组提取模块将报告预处理为三元组。然后，我们对提取出的三元组进行编码；值得强调的是，在编码过程中，我们通过查询医学知识库，将实体转换为详细描述。最后，我们利用三元组改变训练流程，即在实体级别查询基于 Transformer 的融合模块，该模块提供更详细的监督信号。

视觉语言信号的融合模块。在第3.4节，我们介绍使用日常X射线扫描获取的配对图像报告进行训练的流程；在第3.5节，我们介绍推理流程。

3.1. 问题场景

假设我们拥有一个包含 N 个样本的训练集，即 $\mathcal{D}_{\text{train}} = \{(\mathcal{X}_1, \mathcal{T}_1).....,(\mathcal{X}_N, \mathcal{T}_N)\}$，其中 $\mathcal{X}_i$，$\mathcal{T}_i$ 分别指日常扫描中生成的 X射线图像及其对应的医学报告。我们的目标是训练一个视觉语言模型，使其能够诊断是否存在某些疾病，并在空间上定位视觉证据。具体而言，在推理时，我们可以自由地要求系统识别患者患上某种疾病的可能性（无论该疾病在训练期间是否出现）：

$$\hat{s}_i, \hat{m}_i = \Phi_{\text{fusion}}(\Phi_{\text{visual}}(\mathcal{X}_i), \Phi_{\text{textual}}([\text{description}])), \quad (1)$$

其中，$\mathcal{X}_i \in \mathbb{R}^{H\times W\times 3}$ 指来自**测试集**的一个图像样本，H，W 分别表示高度和宽度。$\hat{s}_i \in [0,1]$ 表示根据输入描述推断出的患者患有某种疾病的可能性，$\hat{m}_i \in \mathbb{R}^{H\times W\times 1}$ 表示预测的空间热图，其中在可能提供该疾病视觉依据的像素上具有较高的激活值。在下一节中，我们将介绍带有三元组提取的报告预处理操作。

3.2. 报告预处理

首先，我们提出使用三**元组提取**模块对医学报告进行预处理，通过去除语言语法中不必要的复杂性来简化报告。需要注意的是，后文仅考虑单个采样的图像-报告对（$\mathcal{X}_i$，$\mathcal{T}_i$），并为简洁起见忽略符号中的下标。

我们使用现成的医学命名实体识别（NER）方法 Rad-Graph [29, 65], 对原始报告进行精简，将报告转换为一组三元组，如图2A所示。具体而言，医学关键词可以由 NER 模块提取，并分类为“实体”或“位置”。“实体”指某些临床观察结果，例如“阴影”。“位置”指放射学报告中出现的解剖部位，例如“右肺下叶”。此外，NER 模块还会提供一个“存在”标签，以判断报告中所述实体是存在、缺失还是不确定。基于此，我们可以使用一组三元组即，{实体，位置，存在}，来改写报告中的句子。例如，三元组{阴影，右肺下叶，真} 表示“右肺下叶确实存在阴影”。**请注意**，具有特定“位置”的三元组，其“存在”值并不总是“真”，因为放射科医生可能会指出某些实体在特定位置不存在。

因此，给定一份包含多个句子的报告 $\mathcal{T}$ $\mathcal{T} = \{s_1, s_2, ..., s_M\}$，提取模块会分别处理每个句子，并从报告中构建若干三元组：

$$\Phi_{\text{ex}}(s_j) = \{\text{entity}_n, \text{position}_n, \text{exist}_n\}, n \in [0, t_j], \quad (2)$$

其中，t_j 表示一个句子中包含的实体总数，$n = 0$表示不存在有效实体的特殊情况。经过三元组提取后，每份报告都等同于一组三元组。

讨论。与自然文本不同，医学报告中的信息往往更加精练，放射科医生会指出异常是否存在及其在图像中的位置。同时，医学术语通常较为专业，而在特定词汇范围内（大多列于统一医学语言系统 [5]中），专门设计的命名实体识别方法 [29] 在报告上表现出色。因此，在医学视觉语言预训练中采用三**元组提取**操作，可以避免理解语法所带来的不必要复杂性，同时保留报告中的有用信息。

3.3. 架构

本节将详细介绍我们提出的框架，该框架由三个组件构成，即视觉编码、

知识增强三元组编码和融合模块，如图 2B 和图 2C 所示。

3.3.1 视觉编码

给定一张 X射线图像扫描 $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$，我们可以使用视觉主干网络计算特征：

$$\mathcal{V} = \Phi_{\text{visual}}(\mathcal{X}) \in \mathbb{R}^{h\times w\times d}, \tag{3}$$

h，w，d 分别表示输出的高度、宽度和特征维度。在我们的实验中，采用 ResNet-50 作为视觉主干网络，并取第 4 个残差块的输出。需要注意的是，为了与现有工作进行公平比较，我们做出了这一架构选择 [66, 45, 25, 6], ；而其他视觉主干网络，例如 ViT [17], ，同样可以使用。

3.3.2 知识增强三元组编码

该模块旨在通过融入医学领域知识，对从报告中提取的三元组进行编码，如图 2B 所示。

给定一个三元组 {实体，位置，存在}，对“存在”进行编码很简单，因为它只有三种取值。我们使用 $l \in \{0, 1, -1\}$ 对其进行分词，1 表示 True，0 表示 False，-1 表示不确定。对于“实体”词语，我们通过查询一些易于访问的医学知识库，将其转换为**详细描述** [1][2]，例如，Description([“肺炎”])=“这是一种主要累及肺部的疾病.....表现为阴影和胸腔积液.....” 。尽管方法简单，将实体转换为描述对于实现更可靠的零样本诊断至关重要，因为这能进一步将专业医学实体分解为不同疾病共有的基本属性，从而鼓励模型深入理解视觉证据。对于“位置”词语，我们使用提示“It is located at {position}”来构成句子。最后，我们使用 ClinicalBERT [3] 作为预训练文本编码器，计算“实体”和“位置”的嵌入，然后采用线性多层感知机将嵌入投影到所需维度：

$$e = \Phi_{\text{textual}}(\text{Description}(\{\text{entity}\})) \in \mathbb{R}^{d}, \tag{4}$$

$$p = \Phi_{\text{textual}}(\text{“It is located at \{position\}”}) \in \mathbb{R}^{d'}. \tag{5}$$

每个三元组现在都已嵌入到 $\{e, p, l\}$ 中。

讨论。 提取出的实体是医学术语，只有具备医学背景的读者才能理解；为这些术语补充详细描述，有助于模型深入理解疾病的视觉证据。这类模式可以推广到不同疾病，因为许多属性描述往往是共通的，使模型能够在已见类别上建立隐式关系，并理解未见类别的描述。

[1]维基百科 https://en.wikipedia.org/wiki/[2]统一医学语言系统 [5] https://www.nlm.nih.gov/research/umls/

3.3.3 融合模块

借助报告中的三元组，我们可以**在实体级别**而非整个报告级别监督模型。三元组中的“位置”和“存在”部分可以自然地视为更细粒度的监督标签。具体而言，我们采用基于Transformer的架构，将实体嵌入作为查询，迭代地关注图像嵌入，并输出实体的存在和位置预测结果。

具体而言，我们选取所有训练报告中最常出现的前 $|Q|$ 个实体的嵌入，构成实体查询集 $Q = \{e_1, e_2, ..., e_{|Q|}\}$。实体查询集的详细信息见补充材料（第 A 节）。随后，将 Q 与图像表示 $\mathcal{V}$ 一同输入融合模块进行对齐。融合模块由多个 Transformer 解码器层组成，其中 Q 作为查询，$\mathcal{V}$ 作为键和值。输出随后分别输入两个多层感知机，以推断实体是否存在以及实体的位置：

$$\{\hat{s},\, \hat{p},\, \hat{m}\} = \Phi_{\text{fusion}}(\mathcal{V},\, Q), \tag{6}$$

其中，$\hat{s} \in \mathbb{R}^{|Q|}$ 表示针对每个实体查询的存在性预测，$\hat{p} \in \mathbb{R}^{|Q|\times d'}$ 表示对所有实体的位置预测。**请注意**，$\hat{m} \in \mathbb{R}^{H\times W}$ 表示来自 Transformer 层的交叉注意力图的平均值，并通过最近邻插值上采样至输入图像的尺寸。推理时，$\hat{m}$ 用于定位，因为它自然地充当分割热力图。训练期间，我们不会直接对其计算任何损失。

讨论。 采用 Transformer 解码器能够在图块级计算实体与图像之间的对应关系。因此，图像特征 $\mathcal{V}$ 更适用于下游分割任务，各层交叉注意力图的平均值可直接用于**零样本** 定位，为诊断提供可解释性。此外，Transformer 结构中查询上的默认自注意力也能建立实体之间的关系。

3.4. 训练

给定从配对报告 $\mathcal{T}$ 中提取的一组编码三元组 $\{e, p, l\}$，我们可以根据融合模块的输出计算训练损失。对于存在性预测 $\hat{s}$，我们使用二元交叉熵及其对应的“存在”标签 l；如果 l 是 -1，则直接传递该标签，记作 $\mathcal{L}_{\text{cls}}$。为了监督每个实体查询的位置预测，我们采用对比学习。我们将顶部 $|P|$ 个常见位置嵌入组成位置集 $P = \{p_1, p_2, \cdots, p_{|P|}\}$，从中随机采样 M 个负位置嵌入，并使用三元组中的对应位置嵌入 p 作为正样本：

$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|}\sum_{k=1}^{|Q|}\frac{e^{\langle \hat{p}_k, p_k\rangle}}{e^{\langle \hat{p}_k, p_k\rangle} + \sum_{u=1}^{M} e^{\langle \hat{p}_k, P_{\mathcal{I}(k,u)}\rangle}}, \tag{7}$$

其中，$\langle \cdot, \cdot \rangle$ 表示两个向量的内积，而 $\mathcal{I}(\cdot, \cdot)$ 是一个随机索引采样函数。计算时，位置嵌入未进行归一化。**请注意**，某些实体可能未在报告中提及，因此我们无法在三元组中找到对应的标签。计算损失时，我们会忽略对应的预测结果。

最终损失是两项损失之和：

$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{loc}} + \alpha_2 \mathcal{L}_{\text{cls}}, \tag{8}$$

其中 α_1，α_2 是控制两项损失比例的两个超参数，默认将它们设为 1。0。

讨论。 与现有将图像与整篇报告对齐的方法 [66]相比，我们采用三元组的训练范式，在更细粒度的实体级别提供监督，而非像现有方法那样通常只进行图像与报告之间的全局对齐。

3.5. 推理

在推理时，给定一张测试图像，我们可以直接推断某些实体/疾病是否存在，并定位其视觉证据。具体而言，对于实体查询集 Q中已出现的实体，我们直接采用 Q中的对应元素；对于未出现的实体，则将其替换为用户提供的简短描述，并将其作为额外查询加入实体查询集 Q，**类似于零样本推理**。存在性输出 $\hat{s}$ 可直接用于分类，而目标实体与视觉特征之间的平均交叉注意力 $\hat{m}$ 则用于定位。

4. 实验

本节首先介绍实验所用的数据集，例如预训练数据集和各种下游数据集。然后，我们介绍实现细节和所考虑的基线方法。

4.1. 预训练数据集

MIMIC-CXR v2 [31, 22] 包含超过 22.7 万项图像-报告配对检查数据，数据来自 65,379 名患者在不同扫描时的检查。每项检查包含一张或两张图像（不同扫描视图），总计 377,110 张图像。

4.2. 下游任务的数据集

ChestX-ray14 [56] 包含 112,120 张正位 X 射线图像，来自 30,805 名不同患者，由美国国立卫生研究院（NIH）于 1992 年至 2015年间收集，并提供了 14 种常见疾病的标签。我们将数据集划分为 $0.8/0.1/0.1$，用于训练集/验证集/测试集。

RSNAPneumonia [49] 包含超过 26 万张正位胸部 X 光片，以及相应的肺炎阴影掩膜，这些数据由 RSNA（北美放射学会北美）。通常，该数据集被用于分类任务 [25, 6]。我们将数据集划分为 $0.6/0.2/0.2$，分别用于训练集/验证集/测试集。

SIIM-ACR 气胸 [1] 包含超过 1.2 万张正位胸部 X 光片及其气胸掩码，由 SIIM-ACR（医学影像信息学会和美国放射学会）收集。与 RSNA 肺炎数据集类似，它既可用于分类任务，也可用于分割任务。我们将该数据集划分为$0.6/0.2/0.2$，用于训练集/验证集/测试集。

COVIDxCXR-2 [46]和 COVID Rural [52, 15] 旨在评估新冠肺炎诊断效果。COVIDx CXR-3 包含来自 16,648 名患者的 29,986 张图像，并带有新冠肺炎分类标签。我们将其用作分类数据集，并划分为 $0.7/0.2/0.1$，用于训练集/验证集/测试集。此外，我们使用 COVID Rural 数据集进行新冠肺炎分割。该数据集包含 200 多张带有分割掩码的胸部 X 光片，我们将其划分为 $0.6/0.2/0.2$，用于训练集/验证集/测试集。

水肿严重程度 [7] 包含来自 MIMIC-CXR 的 6,524 个样本，其肺水肿严重程度标签（0 至 3，数值越大表示严重程度越高）是从放射学报告中提取的。其中，141 例由放射科医生评估，并就严重程度等级达成共识。这可视为一项典型的细粒度分类任务。我们将数据集划分为 $0.6/0.2/0.2$ 用于训练集/验证集/测试集。

4.3. 实现

本节介绍各架构的实现。在**预训练**阶段，三元组提取模块以及用于三元组编码的文本编码器均保持固定，而视觉编码器和融合模块则在图像-文本对上进行端到端训练。在**微调**阶段，我们采用以图像编码器初始化的 ResNet50 [24]进行分类，并使用 ResUNet [16]，以我们的预训练图像编码器初始化其编码器来进行分割。有关不同参数的具体值和训练进度的更多详情，请参阅补充材料（第 B 节）。

4.4. 基线方法

我们将本方法与多种现有的先进医学图像-文本预训练方法进行比较，包括 ConVIRT [66],GLoRIA [25], BioViL [6] 和 CheXzero [54]。由于 ConVIRT 和 GLoRIA 是在内部数据集上预训练的，为了公平比较，我们在 MIMIC-CXR 数据集上重新训练了它们的模型。对于 BioViL，我们使用作者正式发布的模型。在零样本设置下，我们采用 BioViL [6] 提及的提示，并与最新方法 CheXzero [54]进行比较；研究表明，该方法的零样本诊断能力优于放射科医生。对于微调，我们均采用第 4.3 节所述的相同设置。

数据集	RSNA 肺炎			SIIM-ACR 气胸			ChestX-ray14		
方法	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [66]	0.8042	0.5842	0.7611	0.6431	0.4329	0.5700	0.6101	0.1628	0.7102
GLoRIA [25]	0.7145	0.4901	0.7129	0.5342	0.3823	0.4047	0.6610	0.1732	0.7700
BioViL [6]	0.8280	0.5833	0.7669	0.7079	0.4855	0.6909	0.6912	0.1931	0.7916
CheXzero [54]	0.8579	0.6211	0.7942	0.6879	0.4704	0.5466	0.7296	0.2141	0.8278
Ours	**0.8694**	**0.6342**	**0.8002**	**0.8924**	**0.6833**	**0.8428**	**0.7676**	**0.2525**	**0.8619**

表1：与零样本分类任务中的其他最先进方法进行比较。报告了曲线下面积、F1分数和准确率分数。对于ChestX-ray14，所有指标均指14种疾病的宏平均。

提示类型	直接新冠肺炎			新冠肺炎描述		
方法	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [66]	0.6159	0.7057	0.6113	0.5208	0.6902	0.5266
GLoRIA [25]	0.6319	0.6938	0.5710	0.6659	0.7007	0.6083
BioViL [6]	0.6137	0.6958	0.5461	0.5382	0.6910	0.5375
CheXzero [54]	0.6462	0.7369	0.6629	0.6667	0.6400	0.6578
Ours	0.6561	0.7066	0.5917	**0.7396**	**0.7670**	**0.7006**

表 2：在零样本新冠肺炎分类任务上与其他先进方法的比较。报告曲线下面积、F1分数和准确率分数。“直接新冠肺炎”指直接使用“新冠肺炎”构建提示句，而“新冠肺炎描述”指用其描述替换“新冠肺炎”这一名称。

4.5. 指标

曲线下面积、F1分数和准确率用于衡量分类任务。F1分数综合衡量模型的召回率和精确率，ACC是准确率的简称。最终的二元预测阈值选择使F1分数最大化的阈值。也在该阈值下计算ACC分数。

指点游戏用于评估定位性能。具体而言，我们提取输出热力图中响应值最大的区域。对于一个实例，如果该区域命中真实掩码，则视为正预测；否则视为负预测。最后，可以将准确率作为指点游戏分数进行计算。

Dice系数和交并比常用于分割任务。对于零样本分割，我们以0.01为间隔搜索所有方法的分割阈值，并报告每个模型的最大Dice系数。

精确率和召回率指检测精确率和召回率。在医学领域，即使无法进行精细分割，检测出病灶也很重要。此外，在一些困难案例中，尤其是在零样本设置下，Dice系数和交并比可能过于严格，无法反映性能差异。精确率和召回率可以弥补这一不足。我们选择交并比阈值0.1来计算这些分数。

5. 结果

本节将报告实验结果。总体而言，我们将结果分为两部分：零样本设置和微调设置。在零样本情况下（第 5.1 节），我们进行消融研究，并将结果与其他最先进的图文预训练方法进行比较。我们主要考虑分类和分割任务；在微调情况下（第 5.2 节），我们通过使用 1%、10% 和 100% 的数据比例对模型进行微调，评估模型的迁移能力。

此外，我们还加入了疾病分级这一项下游任务，它可视为一项细粒度分类任务，表明我们的预训练模型能够轻松迁移至下游任务。

5.1. 零样本评估

本节中，我们在分类和定位任务上，将我们的方法与其他最先进的方法在零样本设置下进行比较。由于篇幅有限，完整的消融研究见补充材料（第 C 节），其中提供了更多细节和分析；此处所有比较均使用我们效果最佳的模型，该模型采用位置对比损失和实体描述编码器。

5.1.1 分类

已见疾病。如表 1 所示，我们在三个广泛使用的数据集上与现有方法进行了比较，结果显示性能持续提升。具体而言，在肺炎和气胸数据集上，尽管图像由不同临床机构采集，且对应不同疾病，我们的模型仍将 RSNA 肺炎数据集上的 AUC 分数从 0.83 提升至 0.87，并将 SIIM-ACR 气胸数据集上的 AUC 分数从 0.71 提升至 0.89。如表 1 所示。这表明，我们的方法能够更好地应对医学领域中的多中心、多疾病数据分布。在 ChestX-ray14 数据集上，我们将平均 AUC 分数从 0.69 提升至 0.77；有关 14 种疾病的详细比较，请参阅补充材料（第 D 节）。

未见疾病。在这里，我们考虑一种严格的开放集分类情形，特别是使用新冠肺炎来评估系统。新冠肺炎是一种直到2019年才出现的新疾病，而2015年收集的 MIMIC-CXR 报告中并不包含任何新冠肺炎实体，因此系统必须具备真正诊断未见疾病的能力

方法	指点游戏↑	召回率↑	精确率↑	IoU↑	Dice系数↑
GLoRIA [25]	0.7607	0.8330	0.1621	0.2182	0.3468
BioViL [6]	0.8342	0.8521	0.5034	0.3029	0.4386
Ours	**0.8721**	**0.8661**	**0.6420**	**0.3172**	**0.4649**

（a）肺炎零样本定位

方法	指点游戏↑	召回率↑	精确率↑
GLoRIA [25]	0.0651	0.2377	0.0585
BioViL [6]	0.0252	0.1963	0.1429
Ours	**0.1975**	**0.3562**	**0.1940**

(b) 气胸零样本定位

表 3：与其他最先进方法在零样本区域定位任务上的比较。(a) 展示了 RSNA肺炎数据集上的结果。(b) 展示了 SIIM-ACR气胸数据集上的结果。气胸区域通常较薄且狭窄，定位难度更高，因此我们仅考虑指点游戏、召回率和精确率。我们的方法在不同指标上均能取得更好的表现，尤其是在指点游戏分数方面。ConVIRT 和 CheXzero 无法实现这一功能。

提示类型	直接新冠肺炎					新冠肺炎描述				
方法	指点游戏↑	召回率↑	精确率↑	IoU↑	Dice系数↑	指点游戏↑	AR↑	AP↑	IoU↑	Dice系数↑
GLoRIA [25]	0.0364	0.2906	0.1073	0.0645	0.1141	0.2727	0.2821	0.1336	0.0596	0.1075
BioViL [6]	0.4000	0.2564	0.2703	0.1198	0.1967	0.1818	0.2393	0.1637	0.0861	0.1427
Ours	0.1818	0.1880	0.1497	0.0747	0.1289	**0.5818**	**0.5214**	**0.4959**	**0.1373**	**0.2278**

表4：零样本新冠肺炎阴影区域定位任务中与其他最先进方法的比较。*"Directcovid-19"* 表示直接使用"新冠肺炎"构造提示句以进行实体编码，而"新冠肺炎描述"表示用其描述替换名称"新冠肺炎"。我们的方法在不同指标上均能取得更好的性能。

如表2所示，仅依赖疾病名称的现有方法难以做出正确诊断。相比之下，我们提出的方法在引入医学知识后，即使用实体描述，能够理解训练集中未见过的复杂医学实体描述，并显著提升性能：AUC 从 0.66 提高到 0.74，准确率从 0.59 提高到 0.70，这表明实体翻译对于未见疾病至关重要。

5.1.2 定位

除了单纯的诊断之外，可解释性在医疗保健领域也同样至关重要，能够提升机器学习系统的可靠性与可信度。在此，我们通过定位预测中的异常来提供可解释性，并将其与现有方法进行比较。同样地，我们根据疾病名称是否出现在医学报告中，将疾病分为已见和未见两类。具体而言，"肺炎"和"气胸"被视为已见疾病，而"新冠肺炎"则被视为未见疾病。受篇幅所限，可视化结果见补充材料（第 E 节）。

已见疾病。 我们在表 3 中展示了 RSNA 肺炎阴影和 SIIM-ACR 气胸肺萎陷定位的结果。如表 3a 所示，我们提出的模型在所有指标上均优于现有方法。例如，我们将指点游戏分数从 0.83 提升至 0.87，将检测召回率从 0.85 提升至 0.87，将检测精确率从 0.50 提升至 0.64，将交并比从 0.30 提升至 0.32，并将 Dice系数从 0.44 提升至 0.46。而在 SIIM-ACR 数据集上（表 3b），由于气胸区域往往细长狭窄，对其进行定位通常比阴影定位更具挑战性 [6], 因此我们仅考察指点游戏、召回率和精确率。类似地，我们的方法也能取得显著优于先前方法的性能。

未见疾病。我们还针对未见疾病（即新冠肺炎）开展了零样本定位实验，如表 4 所示。我们的模型在所有指标上均取得了稳定的提升，例如，指点游戏分数从 0.40 提升至 0.58。值得注意的是，表 4 中的结果与表 2 中的结果大多一致，也就是说，分类结果越好，定位效果往往也越好。总体而言，采用知识增强语言编码的模型使视觉编码器能够学习与疾病相关的潜在证据，从而生成比以往方法更具可解释性的表征。

5.2. 微调评估

在本节中，我们考察微调场景：以预训练模型作为初始化，并针对下游任务进行端到端训练。我们在三种不同的任务上进行测试，分别是分类、分割和分级。在分类和分割任务中，测试集划分和评估指标与"零样本"一节相同。分级是我们在微调设置中引入的一项新任务，可视为细粒度分类任务。

5.2.1 分类

我们在四个不同的数据集上进行实验，使用 1%、10% 和 100% 的数据进行微调，这与现有工作一致 [66, 25, 6]。如表 5 所示，我们的模型

数据集 数据占比	肺炎 1%	肺炎 10%	肺炎 100%	气胸 1%	气胸 10%	气胸 100%	新冠肺炎 1%	新冠肺炎 10%	新冠肺炎 100%	ChestX-ray14 1%	ChestX-ray14 10%	ChestX-ray14 100%
从头训练	0.7107	0.8150	0.8626	0.4347	0.6120	0.6571	0.7861	0.9162	0.9554	0.6005	0.7365	0.7924
ConVIRT [66]	0.8398	0.8562	0.8761	0.7134	0.7826	0.9004	0.8675	0.9541	0.9726	0.6615	0.7658	0.8128
GLoRIA [25]	0.8599	0.8666	0.8846	0.7439	0.8538	0.9014	0.9065	0.9381	0.9728	0.6710	0.7642	0.8184
BioViL [6]	0.8233	0.8538	0.8836	0.6948	0.7775	0.8689	0.8989	0.9529	0.9729	0.6952	0.7527	0.8245
Ours	**0.8731**	**0.8799**	**0.8931**	**0.8527**	**0.9071**	**0.9188**	**0.9224**	**0.9657**	**0.9729**	**0.7721**	**0.7894**	**0.8323**

表 5：在微调分类任务中与其他最先进方法的 AUC分数比较。报告了 ChestX-ray14 数据集上14种疾病的 AUC分数宏平均值。

疾病 数据占比	肺炎 1%	肺炎 10%	肺炎 100%	气胸 1%	气胸 10%	气胸 100%	新冠肺炎 1%	新冠肺炎 10%	新冠肺炎 100%
从头训练	0.4347	0.6047	0.7068	0.2133	0.3323	0.7447	0.1481	0.2367	0.3228
ConVIRT [66]	0.5706	0.6491	0.7201	0.5406	0.6121	0.7352	0.1995	0.2724	0.3737
GLoRIA [25]	0.6555	0.6907	0.7328	0.5673	0.5778	0.7694	0.1889	0.2809	0.3869
BioViL [6]	0.6824	0.7038	0.7249	0.6267	0.6998	0.7849	0.2113	0.3239	0.4162
Ours	**0.7064**	**0.7162**	**0.7579**	**0.6659**	**0.7210**	**0.7937**	**0.2445**	**0.3539**	**0.4399**

表 6：微调分割任务中与其他最先进方法的 Dice 分数比较。报告了三种疾病，并针对每种疾病采用三种数据占比：1%、10% 和 100%，以展示不同数据量下的性能变化。

方法	0 AUC↑	0 F1↑	0 准确率↑	1 曲线下面积↑	1 F1↑	1 准确率↑	2 曲线下面积↑	2 F1↑	2 准确率↑	3 曲线下面积↑	3 F1↑	3 准确率↑	AVG 曲线下面积↑	AVG F1↑	AVG ACC↑
从头训练	0.7631	0.7036	0.6738	0.5383	0.3593	0.3223	0.6692	0.4328	0.7012	0.8420	0.5694	0.8770	0.7031	0.5163	0.6436
ConVIRT [66]	0.8453	**0.7769**	**0.7793**	0.6099	0.3938	0.4629	0.7202	0.4843	0.6445	0.9047	0.6154	0.8809	0.7700	0.5676	0.6919
GLoRIA [25]	0.8304	0.7577	0.7520	0.6208	0.3991	0.4922	0.7339	0.4958	**0.7037**	**0.9246**	**0.6667**	0.9102	0.7774	0.5798	0.7145
BioViL [6]	0.8034	0.7378	0.7148	0.6035	0.3912	0.4570	0.6860	0.4497	0.6777	0.9229	0.6500	0.9160	0.7540	0.5572	0.6914
Ours	**0.8502**	0.7646	0.7539	**0.6641**	**0.4140**	**0.5392**	**0.7605**	**0.5266**	0.7031	0.8845	0.6250	**0.9160**	**0.7898**	**0.5826**	**0.7280**

表7：在水肿严重程度分级多类分类任务微调中与其他最先进方法的比较。表中报告了 AUC 分数。表中的“0,1,2,3”表示严重程度等级，并报告了最终平均分数。

在所有数据集上的曲线下面积分数均显著优于现有方法，表明我们的预训练表示质量高于现有模型。更多详细的比较结果请参阅补充材料（第 D 节）。

5.2.2 分割

在表 6 中，我们针对三种不同疾病进行了分割微调实验。我们选取 1%、10% 和 100% 的数据进行微调。对于图像分布各异的这三种疾病，我们的方法都大幅超越了现有的最先进方法，尤其是在低数据环境下。

5.2.3 分级

除了诊断之外，对疾病严重程度等级进行分级也发挥着重要作用。在此，我们采用预训练特征，并针对多类分类任务对其进行训练，其中 0 至 3 分别代表不同的严重程度等级。如表 7 所示，对于每个等级，均将其作为一类与所有其他类别进行比较，计算曲线下面积、F1分数和准确率分数，例如，0 与 $\{1,2,3\}$。随后计算四个等级的最终宏平均分数。在大多数严重程度等级上，我们的方法都能取得最佳结果。

6. 结论

本文介绍了一种新颖的医学知识增强型视觉语言预训练模型。首先，我们提出一个三元组提取模块，用于提取有用的医学相关三元组，作为更有效的监督信号，从而在尽可能减少信息损失的同时简化复杂的原始报告。其次，我们将提取出的三元组中的实体转换为详细的医学描述，并通过文本编码器对其进行嵌入，使网络能够理解复杂的专家级医学知识。最后，我们提出一种基于Transformer的结构来实现局部区域对齐。在实验中，我们在不同数据集和多种设置下评估了所提方法。即使面对未见疾病，我们的方法也展现出出色的零样本分类和定位能力。此外，经过微调后，我们的方法仍显著优于最先进的方法，体现了其优越性。

参考文献

[1] 医学影像信息学会：SIIM-ACR气胸分割。https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation. 2019年。[2] Emily Alsentzer、John Murphy、William Boag、Wei-Hung Weng、Di Jin、Tristan Naumann和Matthew McDermott。公开可用的临床 BERT 嵌入。载于第*2*届临床自然语言处理研讨会论文集，第72–78页，美国明尼苏达州明尼阿波利斯，2019年6月。计算语言学协会。[3]Emily Alsentzer、John R Murphy、Willie Boag、Wei-Hung Weng、Di Jin、Tristan Naumann、WA Redmond和Matthew BA McDermott。公开可用的临床BERT嵌入。*NAACLHLT2019*，第72页，2019年。[4] Federico Bianchi、Giuseppe Attanasio、Raphael Pisoni、Silvia Terragni、Gabriele Sarti和Sri Lakshmi。意大利语的对比式语言-图像预训练。*arXiv*预印本 *arXiv:2108.08688*，2021年。[5] Olivier Bodenreider。统一医学语言系统（UMLS）：整合生物医学术语。《核酸研究》，32(suppl 1):D267–D270，2004年。[6]Benedikt Boecking、Naoto Usuyama、Shruthi Bannur、Daniel C Castro、Anton Schwaighofer、Stephanie Hyland、Maria Wetscherek、Tristan Naumann、Aditya Nori、Javier Alvarez-Valle等。充分利用文本语义以改进生物医学视觉语言处理。载于欧洲计算机视觉会议，第1–21页，2022年。官方实现：https://github.com/microsoft/hi-ml/tree/main/hi-ml-multimodal。[7] Geeticka Chauhan、Ruizhi Liao、William Wells、Jacob Andreas、Xin Wang、Seth Berkowitz、Steven Horng、Peter Szolovits和Polina Golland。联合建模胸部X光片和放射学报告以评估肺水肿。载于国际医学图像计算与计算机辅助干预会议，第529–539页。施普林格，2020年。[8] Sihong Chen、Jing Qin、Xing Ji、Baiying Lei、Tianfu Wang、Dong Ni和Jie-Zhi Cheng。利用多任务特征自动评分多个语义属性：一项关于CT图像中肺结节的研究。*IEEE* 医学影像汇刊，36(3):802–814，2016年。[9] Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng和Jingjing Liu。UNITER：通用图像-文本表征学习。载于欧洲计算机视觉会议，第104–120页。施普林格，2020年。[10] Zhihong Chen、Guanbin Li和Xiang Wan。Align、推理与学习：利用知识增强医学视觉语言预训练。载于第*30*届*ACM*国际多媒体会议论文集，第5152–5161页，2022年。[11] Zhihong Chen、Yan Song、Tsung-Hui Chang和Xiang Wan。通过记忆驱动的Transformer生成放射学报告。载于*2020*年实证研究会议论文集自然语言处理方法（*EMNLP*），第 1439–1449 页，2020年。[12] Hui Cui、Yiyue Xu、Wanlong Li、Linlin Wang 和 Henry Duh。从 CT 影像预测放射治疗相关食管瘘的跨通道临床注意力协同学习。载于国际医学图像计算与计算机辅助干预会议，第 212–220 页。施普林格，2020年。[13] Yuhao Cui、Zhou Yu、Chunqi Wang、Zhongzhou Zhao、Ji Zhang、Meng Wang 和 Jun Yu。Rosita：通过跨模态与模态内知识整合，增强视觉与语言语义对齐。载于第 *29* 届 *ACM* 国际多媒体会议论文集，第 797–806 页，2021年。[14] Jean-Benoit Delbrouck、Pierre Chambon、Christian Bluethgen、Emily Tsai、Omar Almusa 和 Curtis P Langlotz。通过语义奖励提高放射学报告生成的事实准确性。*arXiv* 预印本 *arXiv:2210.12186*，2022年。[15] Shivang Desai、Ahmad Baghal、Thidathip Wongsurawat、Piroon Jenjaroenpun、Thomas Powell、Shaymaa Al-Shukri、Kim Gates、Phillip Farmer、Michael Rutherford、Geri Blake 等。代表美国农村新冠肺炎阳性人群的胸部影像。*Scientific Data*，7(1):1–6，2020年。[16] Foivos I Diakogiannis、Franc¸ois Waldner、Peter Caccetta 和 Chen Wu。ResUNet-a：用于遥感数据语义分割的深度学习框架。*ISPRS*摄影测量与遥感期刊，162:94–114，2020年。[17] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于国际学习表征会议，2020年。[18] Jared A Dunnmon、Alexander J Ratner、Khaled Saab、Nishith Khandwala、Matthew Markert、Hersh Sagreiya、Roger Goldman、Christopher Lee-Messer、Matthew P Lungren、Daniel L Rubin 等。跨模态数据编程助力快速开展医学机器学习。*Patterns*，1(2):100019，2020年。[19] Bradley J Erickson、Panagiotis Korfiatis、Zeynettin Akkus 和 Timothy L Kline。医学影像中的机器学习。*Radiographics*，37(2):505，2017年。[20] Andre Esteva、Brett Kuprel、Roberto A Novoa、Justin Ko、Susan M Swetter、Helen M Blau 和 Sebastian Thrun。利用深度神经网络实现皮肤癌的皮肤科医生级分类。《自然》，542(7639):115–118，2017年。[21] Leyuan Fang、Chong Wang、Shutao Li、Hossein Rabbani、Xiangdong Chen 和 Zhimin Liu。关注病变：用于视网膜光学相干断层扫描图像分类的病变感知卷积神经网络。*IEEE* 医学影像汇刊，38(8):1959–1970，2019年。[22] Ary L Goldberger、Luis AN Amaral、Leon Glass、Jeffrey M Hausdorff、Plamen Ch Ivanov、Roger G Mark、Joseph E Mietus、George B Moody、Chung-Kang Peng 和 H Eugene Stanley。PhysioBank、PhysioToolkit 和 PhysioNet：复杂生理信号研究新资源的组成部分。《循环》，101(23):e215–e220，2000年。

[23] Ivan Gonzalez-Diaz。Dermaknet：将皮肤科医生的知识融入用于皮肤病变诊断的卷积神经网络。*IEEE* 生物医学与健康信息学期刊，23(2):547–559，2018年。[24] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。[25] Shih-Cheng Huang、Liyue Shen、Matthew P Lungren 和 Serena Yeung。GLoRIA：一种用于标签高效医学图像识别的多模态全局-局部表征学习框架。载于*IEEE/CVF*国际计算机视觉会议论文集，第 3942–3951 页，2021年。官方实现：https://github.com/marshuang80/gloria。[26] Xin Huang、Yu Fang、Mingming Lu、Fengqi Yan、Jun Yang 和 Yilu Xu。Dual-ray net：利用正位和侧位胸部X线片自动诊断胸部疾病。医学影像与健康信息学杂志，10(2):348–355，2020年。[27] Sarfaraz Hussein、Kunlin Cao、Qi Song 和 Ulas Bagci。基于 3D CNN 多任务学习的肺结节风险分层。载于医学影像信息处理国际会议，第 249–260 页。施普林格，2017年。[28] Jeremy Irvin、Pranav Rajpurkar、Michael Ko、Yifan Yu、Silviana Ciurea-Ilcus、Chris Chute、Henrik Marklund、Behzad Haghgoo、Robyn Ball、Katie Shpanskaya 等。CheXpert：包含不确定性标签并支持专家比较的大型胸部X线影像数据集。载于*AAAI* 人工智能大会论文集，第 33 卷，第 590–597 页，2019年。[29]Saahil Jain、Ashwin Agrawal、Adriel Saporta、Steven Truong、Tan Bui、Pierre Chambon、Yuhao Zhang、Matthew P Lungren、Andrew Y Ng、Curtis Langlotz 等。RadGraph：从放射学报告中提取临床实体及其关系。载于神经信息处理系统第三十五届会议数据集与基准赛道（第一轮），2021年。[30] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉-语言表征学习。载于国际机器学习会议，第 4904–4916 页。PMLR，2021年。[31] AEWP Johnson、Tom Pollard、Roger Mark、Seth Berkowitz 和 Steven Horng。MIMIC-CXR 数据库。*PhysioNet10*，13026，2019年。[32] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：通过自举式语言-图像预训练实现统一的视觉-语言理解与生成。*arXiv* 预印本 *arXiv:2201.12086*，2022年。[33] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。先对齐后融合：结合动量蒸馏的视觉与语言表征学习。神经信息处理系统进展，34:9694–9705，2021年。[34] Liu Li、Mai Xu、Xiaofei Wang、Lai Jiang 和 Hanruo Liu。基于注意力的青光眼检测：一个大规模数据库以及卷积神经网络模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10571–10580页，2019年。[35] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。Visualbert：一种简洁且高效的视觉与语言基线模型。*arXiv*预印本 *arXiv:1908.03557*，2019年。[36] Xiaomeng Li、Xiaowei Hu、Lequan Yu、Lei Zhu、Chi-Wing Fu 和 Pheng-Ann Heng。CANet：用于糖尿病视网膜病变和糖尿病黄斑水肿联合分级的跨疾病注意力网络。*IEEE* 医学影像汇刊，39(5):1483–1493，2019年。[37] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei 等。Oscar：面向视觉语言任务的对象语义对齐预训练。载于计算机视觉——*ECCV 2020*：*16 28*，*2020*，*23 16*，第121–137页届欧洲会议，英国格拉斯哥，*8*月——论文集，第*XXX*部分 。施普林格，2020年。[38] Qing Liao、Ye Ding、Zoe L Jiang、Xuan Wang、Chunkai Zhang 和 Qian Zhang。用于癌症诊断的多任务深度卷积神经网络。神经计算，348:66–73，2019年。[39] Fenglin Liu、Xian Wu、Shen Ge、Wei Fan 和 Yuexian Zou。探索并提炼用于放射学报告生成的后验与先验知识。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第13753–13762页，2021年。[40] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于国际学习表征会议，2018年。[41] Gabriel Maicas、Andrew P Bradley、Jacinto C Nascimento、Ian Reid 和 Gustavo Carneiro。像放射科医生一样训练医学图像分析系统。载于医学图像计算与计算机辅助干预国际会议，第546–554页。施普林格，2018年。[42] Matthew BA McDermott、Tzu Ming Harry Hsu、Wei-Hung Weng、Marzyeh Ghassemi 和 Peter Szolovits。CheXpert++：通过近似CheXpert标注器，实现速度、可微性和概率输出。载于医疗保健机器学习会议，第913–927页。PMLR，2020年。[43]Masahiro Mitsuhara、Hiroshi Fukui、Yusuke Sakashita、Takanori Ogata、Tsubasa Hirakawa、Takayoshi Yamashita 和 Hironobu Fujiyoshi。通过注意力图将人类知识嵌入深度神经网络。*arXiv*预印本 *arXiv:1905.03540*，2019年。[44] Yasuhide Miura、Yuhao Zhang、Emily Tsai、Curtis Langlotz 和 Dan Jurafsky。提升图像到文本的放射学报告生成的事实完整性与一致性。载于*2021*年计算语言学协会北美分会人类语言技术会议论文集，第5288–5304页，2021年。[45] Philip M¨uller、Georgios Kaissis、Congyu Zou 和 Daniel R¨uckert。从医学图像和报告中联合学习局部化表征。*arXiv*预印本 *arXiv:2112.02889*，2021年。[46] Maya Pavlova、Naomi Terhljan、Audrey G Chung、Andy Zhao、Siddharth Surana、Hossein Aboutalebi、Hayden Gun-

raj、Ali Sabri、Amer Alaref 和 Alexander Wong。Covid-net cxr-2：一种增强型深度卷积神经网络设计，用于检测胸部X光图像中的新冠肺炎病例。*Frontiers in Medicine*，第9卷，2022年。[47] Yifan Peng、Xiaosong Wang、Le Lu、Mohammadhadi Bagheri、Ronald Summers 和 Zhiyong Lu。Negbio：一种用于放射学报告中否定与不确定性检测的高性能工具。*AMIA*转化科学峰会论文集，2018:188，2018年。[48]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页。PMLR，2021年。[49]George Shih、Carol C Wu、Safwan S Halabi、Marc D Kohli、Luciano M Prevedello、Tessa S Cook、Arjun Sharma、Judith K Amorosa、Veronica Arteaga、Maya Galperin-Aizenberg 等。通过可能肺炎的专家标注扩充美国国立卫生研究院胸部X线影像数据集。放射学：人工智能，1(1)，2019年。[50] Akshay Smit、Saahil Jain、Pranav Rajpurkar、Anuj Pareek、Andrew Y Ng 和 Matthew Lungren。结合自动标注器与专家标注，使用BERT准确标注放射学报告。载于*2020*年计算语言学经验方法会议论文集（*EMNLP*），第1500–1519页，2020年。[51] Jiaxing Tan、Yumei Huo、Zhengrong Liang 和 Lihong Li。融合专家知识的深度学习用于肺结节自动检测。《*X*射线科学与技术杂志》，27(1):17–35，2019年。[52] Haiming Tang、Nanfei Sun 和 Yi Li。用于自动检测新冠阳性患者胸部X光片阴影区域并评估疾病严重程度的深度学习分割模型。*medRxiv*预印本，2020年。[53] Joseph J Titano、Marcus Badgeley、Javin Schefflein、Margaret Pain、Andres Su、Michael Cai、Nathaniel Swinburne、John Zech、Jun Kim、Joshua Bederson 等。利用深度神经网络自动监测颅脑影像中的急性神经系统事件。《自然·医学》，24(9):1337–1341，2018年。[54] Ekin Tiu、Ellie Talius、Pujan Patel、Curtis P Langlotz、Andrew Y Ng 和 Pranav Rajpurkar。通过自监督学习，从未标注胸部X光图像中实现专家级病理状况检测。《自然·生物医学工程》，第1–8页，2022年。[55] Kun Wang、Xiaohong Zhang、Sheng Huang、Feiyu Chen、Xiangbo Zhang 和 Luwen Huangfu。利用知识引导的深度缩放神经网络学习从胸部X光片中识别胸部疾病。*IEEE Access*，8:159790–159805，2020年。[56] Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu、Mohammadhadi Bagheri 和 Ronald M Summers。ChestX-ray8：医院规模的胸部X光数据库，以及常见胸部疾病弱监督分类与定位的基准。载于*IEEE*计算机视觉与模式识别会议论文集，第2097–2106页，2017年。[57] Zhanyu Wang、Mingkang Tang、Lei Wang、Xiu Li 和 Lup- ing Zhou。医学语义辅助 Transformer 用于放射学报告生成。载于 医学图像计算与计算机辅助干预–*MICCAI 2022*：第 *25* 届国际会议，新加坡，*2022*年*9*月*18*日至*22*日，会议录，第三部分，第 655–664 页。施普林格，2022年。[58] Zifeng Wang、Zhenbang Wu、Dinesh Agarwal 和 Jimeng Sun。Medclip：从未配对的医学图像和文本中进行对比学习。*arXiv* 预印本 *arXiv:2210.10163*，2022年。[59] Joy T Wu、Nkechinyere Nneka Agu、Ismini Lourentzou、Arjun Sharma、Joseph Alexander Paguio、Jasper Seth Yao、Edward Christopher Dee、William G Mitchell、Satyananda Kashyap、Andrea Giovannini 等。用于临床推理的 Chest ImaGenome 数据集。载于 第三十五届神经信息处理系统大会数据集与基准赛道（第二轮），2021年。[60] Xiaozheng Xie、Jianwei Niu、Xuefeng Liu、Zhengsu Chen、Shaojie Tang 和 Shui Yu。将领域知识融入医学图像分析深度学习的综述。*MedicalImage Analysis*，69:101985，2021年。[61] Yutong Xie、Yong Xia、Jianpeng Zhang、Yang Song、Dagan Feng、Michael Fulham 和 Weidong Cai。基于知识的协作式深度学习用于胸部 CT 良恶性肺结节分类。*IEEE* 医学影像汇刊，38(4):991–1004，2018年。[62] Shuxin Yang、Xian Wu、Shen Ge、S Kevin Zhou 和 Li Xiao。知识至关重要：利用通用知识与特定知识生成胸部放射学报告。*Medical Image Analysis*，80:102510，2022年。[63] Wenkai Yang、Juanjuan Zhao、Yan Qiang、Xiaotang Yang、Yunyun Dong、Qianqian Du、Guohua Shi 和 Muhammad Bilal Zia。DSCGANs：在训练双路径半监督条件生成对抗网络和 S3VM 进行超声检查甲状腺结节分类时融入领域知识。载于 医学图像计算与计算机辅助干预国际会议，第 558–566 页。施普林格，2019年。[64] Fei Yu、Jiji Tang、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。ERNIE-ViL：通过场景图增强知识的视觉-语言表示。载于 *AAAI* 人工智能大会论文集，第 35 卷，第 3208–3216 页，2021年。[65] Ke Yu、Shantanu Ghosh、Zhexiong Liu、Christopher Deible 和 Kayhan Batmanghelich。胸部 X 光片中解剖结构引导的弱监督异常定位。载于 医学图像计算与计算机辅助干预国际会议，第 658–668 页。施普林格，2022年。[66] Yuhao Zhang、Hang Jiang、Yasuhide Miura、Christopher D Manning 和 Curtis P Langlotz。从配对图像和文本中对医学视觉表示进行对比学习。载于 医疗保健机器学习，2022年。最高星级实现：https://github.com/ edreisMD/ConVIRT-pytorch。

A. 实体描述基准和位置集

我们利用 [65] 提供的 rad-graph 命名实体识别提取结果，并进一步添加额外描述。表8展示了我们用于翻译不同实体的描述。按照 [65], , 我们在实体查询集 Q中保留了 75 个实体，该集合覆盖报告中的 90% 实体。“尾部异常观察”实体表示一些尾部实体，而“排除观察”表示一些对诊断无用的实体。最后一行“新冠肺炎”仅用于推理，因为它从未出现在预训练报告中。

表 8：实体描述　　用于翻译单个实体名称的描述。该描述可轻松从网站找到　　.

实体	描述
正常	这表示没有疾病或病症，说明该结构正常。
清晰	肺部清晰且正常。肺部未见其他疾病的证据。
锐利	这表示解剖结构的边界或边缘清晰且正常，也就是说没有疾病。
清晰地	“清晰可见”是指某个解剖结构清楚可见。
未见异常	这表示某些解剖结构正常，通常用于描述心影和纵隔影。
完整	骨性结构完整且正常，表示没有骨折。
稳定	改动后的解剖结构正常且稳定。未发现疾病迹象。
free	通常指游离空气，与气胸、肺不张、气腹和肺气肿有关。
积液	胸腔积液是过量液体积聚在胸膜腔内所致。胸膜腔是环绕每侧肺的潜在腔隙。胸腔积液浸润脏层胸膜与壁层胸膜之间的腔隙。
阴影	它被定义为因液体导致空气移位、气道塌陷、纤维化或肿瘤性病变而形成的模糊阴影区域。它病因包括感染、间质性肺病和肺水肿。
气胸	气胸是指肺与胸壁之间的胸膜腔内异常积聚空气。其可能由以下因素引起：肺炎或纤维化及其他疾病。
水肿	肺水肿，也称为肺充血，是肺组织和含气腔隙中液体过度积聚的现象。它会显示肺泡壁中存在液体。
肺不张	它是肺部塌陷或闭合，导致气体交换减少或缺失。检查发现可能包括肺阴影化和肺容积减少。
tube	这是一种外科引流管，经胸壁插入胸膜腔或纵隔，以移除不需要的物质例如空气。
实变	这是一个通常可压缩的肺组织区域，其中充满了液体而非空气。诊断必须存在实变肺炎：大叶性肺炎的征象具有特征性，临床上称为实变。
病变	急性病变意味着解剖结构存在异常。
异常	它表示疾病和病症的存在，说明结构异常。
扩大	它通常用于修饰心影和心脏。心脏扩大是一种心脏增大的医学状况。
tip	它指的是导管的顶端。
low	肺容积减少可能是限制性肺疾病的征兆，例如肺纤维化或结节病。
肺炎	肺炎是一种肺部炎症，主要累及称为肺泡的小型肺气囊。肺炎可能表现为阴影。也可能发现胸腔积液等并发症，从而提高肺实变和胸腔积液的诊断准确性
line	它指的是静脉通路或经外周静脉置入中心静脉导管。
充血	肺充血是指液体积聚在肺内，导致气体交换受损和动脉低氧血症。
导管	导管是一种置入体内、用于引流和收集膀胱尿液的管子。
心脏扩大	心脏扩大（有时也称为心脏肥大）是一种心脏增大的医学状况。
骨折	骨折是肋骨断裂。
air	指胸膜腔内存在游离空气或气体，表明有气胸。液体导致的空气移位可能会造成阴影。
迂曲	主动脉轻度迂曲。有时也可能指静脉曲张
lead	指导管的前端。
疾病	表示存在疾病和异常，说明结构异常。
钙化	肺钙化是一种常见的无症状检查发现。肺钙化主要由两种机制引起：营养不良型和转移型
显著性	这意味着存在某种观察结果。
医疗器械	它指的是一些设备，例如PICC管、瓣膜导管、心脏起搏器硬件、关节置换标记物、植入式心律转复除颤器、器械支持设备以及输液港
充血	肺血管充血是指肺部血管网络中的血液正常流动受阻，导致肺血管充血
picc	经外周静脉置入中心静脉导管（PICC），也称为PICC导管，是一根细长的导管，经手臂上的静脉插入并延伸至靠近心脏的大静脉。
clip	外科夹或血管夹通常属于医疗设备。
抬高	如果组织或解剖结构抬高了，就表示它们的位置高于正常位置。
扩张	这表示肺正常扩张且清晰，说明没有气胸。
结节	肺结节或肺部结节是肺部相对较小的局灶性密度，可能会与某种结构的投影混淆例如胸壁或皮肤上的结构，如乳头、愈合中的肋骨骨折或肺癌。
wire	胸骨切开术钢丝是指胸部的中心线。
液体	它指肺部的水分或液体，可能表明水肿及其他疾病。
退行性	退行性疾病是基于细胞退行性改变持续发展的结果
起搏器	起搏器设备通常代表一种医疗设备。

实体	描述
增厚	胸膜增厚是指一侧或双侧肺胸膜变厚。其可能由肺部感染引起，脓胸、结核病或肺癌。
标记	它表示间质纹理或支气管血管纹理
scar	瘢痕（或瘢痕组织）是损伤后取代正常组织的纤维组织区域。
过度充气	肺过度充气是指潴留气体导致肺部体积大于正常。
变钝	肋膈角变钝通常由胸腔积液引起，正如前文所述。肋膈角变钝的其他原因肋膈角区域的肺部疾病，以及肺过度充气，均可导致肋膈角变钝。
loss	肺容积减少的病因包括：气道阻塞或压迫、肥胖、脊柱侧凸、限制性疾病例如肺纤维化和间质性肺病、结核病。
增宽	纵隔未见增宽或增大。
肺萎陷	肺萎陷指的是气胸或肺不张。
密度	物质的密度（更准确地说是体积质量密度，也称为比质量）是其单位体积的质量。
肺气肿	肺气肿（或肺气肿症）是一种下呼吸道疾病，其特征是肺内出现含气腔隙（肺气肿），这些腔隙大小不一，有时可能非常大。
充气	通气（也称为曝气或充气）是空气在液体或其他起流体作用的物质（如土壤）中循环、混合或溶解的过程。
mass	肺部肿块是肺内的异常生长或区域，也可视为肺癌。
拥挤	支气管血管结构拥挤是容积减少的重要直接征象。肺不张在增强后呈明显强化由于萎陷肺叶内肺动脉和小动脉彼此靠近而进行对比剂给药。
浸润	肺部浸润是指一种密度高于空气的物质，例如脓液、血液或蛋白质，其滞留在肺实质内。肺部浸润与肺炎、结核病和结节病有关。
模糊不清	一些解剖结构不清晰，难以理解或观察。
畸形	指某些身体部位异常或受伤。
疝	肺疝（西布森疝）是肺组织突出至胸壁外的一种情况。疝可能发生在胸部创伤、胸外科手术或某些肺部疾病之后。
引流	管式引流是一种医疗设备。
膨胀	膨胀通常指增大、扩张或鼓胀效应。它可能指：腹胀。
移位	纵隔移位是指纵隔结构向胸腔一侧偏移，通常可在胸部X线片上观察到。这表明胸腔内压力严重不对称。
支架	气管支架是一种医疗设备
压力	在所有生理压力中，肺静脉压介于平均肺动脉压和左心房压之间
病变	肺结节、肺部结节、白斑、病灶——这些术语都描述了同一种现象：肺部异常。
检查发现	对身体部位的某种观察，通常表示存在异常。
临界	心影大小处于临界范围，意味着心影未增大，且属正常。
医疗器械	它代表一种医疗设备。
扩张	比正常状态更大或更宽的状态
chf	心力衰竭（有时称为充血性心力衰竭）发生在心肌泵血能力不如正常时。发生这种情况时，血液通常会淤积，液体可能积聚在肺部，从而导致呼吸急促。
再分布	如果肺水肿是由心力衰竭或液体过负荷引起的，还可能会看到心脏扩大和肺静脉扩张，尤其是在上肺野。
误吸	吸入性肺炎发生在食物或液体被吸入气道或肺部，而不是被吞咽时。
尾部异常观察	一些非常罕见的疾病。
排除观察	一些毫无意义的观察结果。
新冠肺炎	这是一种由病毒引起的传染病。磨玻璃影、实变、增厚和胸腔积液常见于感染。

此外，我们保留了51 个阳性位置，按照 [65], 构成位置集 P，包括 {气管、左肺门、右肺门、未指明肺门、左侧胸膜、右侧胸膜、未指明胸膜、心脏大小、心缘、左侧膈肌、右侧膈肌、未指定膈肌、心后区、左下叶、左上叶、右下叶、右中叶、右上叶、左下肺、左中肺、左上肺、左肺尖、未指明左肺、右下肺、右中肺、右上肺、右肺尖、未指明右肺、肺尖、肺底、左侧肋膈角、右侧肋膈角、未指明肋膈角、心膈角、纵隔、脊柱、锁骨、肋骨、胃、右心房、右心室、主动脉、上腔静脉、间质、实质、腔静脉－心房交界处、心肺、肺、肺容积、未指定、其他}。“其他”用于表示一些尾部位置。

B. 实现细节

模型架构。 作为模型的输入，图像会被调整为 $224 \times 224 \times 3$。我们使用 ResNet50 [24] 的前四层作为视觉主干网络（Φ_{visual}），并采用一个多层感知机层将输出特征维度转换为$d = 256$。因此，视觉编码器输出的特征图为 $\mathcal{V} \in \mathbb{R}^{14\times14\times256}$。在报告端，我们使用预训练命名实体识别模块提取三元组，如 [29], 所述，并计算实体和位置嵌入，其采用

预训练的 ClinicalBERT [2], , 其默认嵌入维度为 $d' = 768$。按照 [65], 我们获得报告中最常出现的 $|Q| = 75$ 个实体和 $|P| = 51$ 个位置。我们为每个实体采样 $M = 7$ 个负位置，以计算对比损失。在融合模块中，我们采用包含 4 个头的 4 层 Transformer 解码器。

预训练。 在此阶段，预处理操作和语言编码均使用预训练网络，而视觉编码器和融合模块则进行端到端训练。我们使用 AdamW [40] 优化器，设置为 $lr = 1 \times 10^{-4}$ 和 $lr_{\text{warm up}} = 1 \times 10^{-5}$。我们在一块 GeForce RTX 3090 图形处理器上训练，批量大小为 32，共训练 60 个训练轮次。前 5 个训练轮次用于预热。

微调。 对于下游任务，在训练数据量较大的情况下，我们可以使用预训练视觉主干网络进行初始化，并对模型进行端到端微调。具体而言，对于图像分类任务，我们采用 ResNet50 [24] , 并使用预训练视觉编码器初始化其前四层。对于图像分割任务，我们使用 ResUNet [16] 作为主干网络，并使用预训练图像编码器初始化其编码器。

C. 消融研究

我们最终的方法主要包含三个关键部分：基于 Transformer 的实体查询融合模块、位置对比损失（PosCL）和实体翻译（ET）编码。我们逐步移除这些模块，以分析其有效性。“w/o (ET)”表示移除 ET 模块，“w/o (PosCL + ET)”表示仅保留带有基础交叉熵损失的融合模块。基于 Transformer 的实体查询融合模块是支撑预训练的最基本模块，因此无法将其移除。表 9 和表 10 展示了定量结果。

实体查询融合模块。 表格中关于“w/o (PosCL +ET)”的行展示了仅采用基础实体存在交叉熵损失进行修改后的基础模型性能。该模型的表现优于许多以往方法。这证明了我们的假设：复杂的句法会显著影响网络捕获有用实体，而我们的预处理操作和实体级监督能够大幅缓解这一问题。

位置对比损失。 PosCL 能显著帮助网络定位异常。如添加 PosCL 后的结果所示，分类结果可进一步提升，例如，在 ChestX-ray14 数据集上的曲线下面积从 0.75 提升至 0.76。除了分类之外，位置对比损失还能进一步提升定位效果。这些结果表明，位置是报告中的另一个重要要素，尤其对于定位任务而言。我们提取的三元组能够在几乎不损失信息的情况下对报告进行归纳和精简，并使网络能够更直接地学习报告信息。

实体翻译。 通过加入实体翻译，我们希望实现两个目标。首先，除了从图像-报告数据中学习之外，网络还能够根据实体描述主动学习不同实体之间的关系。如表格所示，在大多数场景中加入描述有助于网络更好地理解实体，并提升最终指标得分。其次，更重要的是，实体翻译使我们的模型能够**处理开放集中的新疾病**。如果排除 TE，并像以往工作那样仅使用实体名称作为提示，那么在面对未见疾病时，我们方法的性能将显著下降；正文中关于新冠肺炎零样本分类的讨论也说明了这一点。

数据集	RSNA 肺炎			SIIM-ACR 气胸			ChestX-ray14		
方法	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
不含（PosCL +ET）	0.8532	0.6079	0.7669	0.8768	0.6672	0.8187	0.7502	0.2374	0.8541
不含（ET）	0.8537	0.6241	**0.8146**	**0.9017**	**0.7008**	**0.8584**	0.7621	0.2452	0.8606
Ours	**0.8694**	**0.6342**	0.8002	0.8924	0.6833	0.8428	**0.7676**	**0.2525**	**0.8619**

表9：零样本分类任务的消融研究。报告了曲线下面积、F1分数和准确率分数。对于 ChestX-ray 14，这些指标均指14种疾病的宏平均。

方法	指点游戏↑	召回率↑	精确率↑	IoU↑	Dice系数↑
不含（PosCL + ET）	0.7979	**0.8961**	0.4036	0.2783	0.4230
不含（ET）	0.8424	0.8226	**0.6520**	0.3118	0.4610
Ours	**0.8721**	0.8661	0.6420	**0.3172**	**0.4649**

（a）肺炎零样本定位

方法	指点游戏↑	召回率↑	精确率↑
去除（PosCL + ET）	0.1786	0.3151	0.1336
去除（ET）	**0.2080**	0.3178	0.1711
Ours	0.1975	**0.3562**	**0.1940**

（b）气胸的零样本定位

表10：零样本定位任务上的消融研究。(a) 展示了 RSNA 肺炎数据集上的结果。(b) 展示了 SIIM-ACR气胸数据集上的结果。

D. ChestX-ray14 上的详细结果

我们进一步展示了 ChestX-ray14 数据集上 14 种不同疾病的详细性能。表 11 展示了零样本设置下的结果。对于大多数疾病，我们的方法都能超过以往的方法。雷达图 3Y 更直观地展示了在零样本设置下我们的模型与其他解决方案的比较结果。对于大多数疾病，我们的方法都能超过以往的方法。在 100% 微调设置下，我们同样取得了出色的结果，如表 12 所示。

方法	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
ConVIRT [66]	0.4533	0.4601	0.7262	0.6238	0.6790	0.6322	0.6097	0.6698	0.6855	0.7699	0.4701	0.5293	0.6098	0.6220	0.6101
GLoRIA [25]	0.6680	0.7647	0.7975	0.6159	0.6722	0.5293	0.6755	0.4785	0.7306	0.8212	0.6033	0.5104	**0.6721**	0.7144	0.6610
BioViL [6]	0.5026	0.6328	0.7914	0.5791	0.7029	0.6126	0.6866	0.7516	0.7455	0.8533	0.7136	0.6751	0.6560	0.7692	0.6909
CheXzero [54]	0.7426	0.7956	0.8415	0.6223	0.7095	0.6666	0.7263	0.7679	0.7866	0.8862	0.6451	0.6402	0.6134	0.7704	0.7296
不含（PosCL +ET）不使用（ET）	0.7131	0.8100	0.8635	**0.6361**	0.7776	0.6740	0.6903	0.8124	0.7915	0.8869	0.7480	**0.6780**	0.6429	0.7784	0.7502
	0.7420	0.8270	**0.8663**	0.6336	0.7867	**0.6974**	**0.7238**	0.8310	0.8037	0.8887	0.7865	0.6715	0.5414	0.8691	0.7621
ours	**0.7506**	**0.8299**	0.8636	0.6280	**0.7885**	0.6947	0.7236	**0.8361**	**0.8079**	**0.8888**	**0.7950**	0.6511	0.5783	**0.9097**	**0.7676**

表 11：与其他最先进方法在零样本 ChestX-ray 14 种疾病分类任务上的比较。报告了每种疾病的 AUC 分数，同时也报告了宏平均 AUC 分数。我们使用前三个字母表示一种疾病，但对于“pneumonia”和“pneumothorax”，则使用前两个字母和最后一个字母。

方法	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pna.	Pnx.	Con.	Ede.	Emp.	Fib.	Thi.	Her.	AVG
从头训练	0.7835	0.8116	0.8563	0.6537	0.7788	0.6912	0.7004	0.8561	0.8090	0.8869	0.8564	0.7534	0.7454	0.9106	0.7924
ConVIRT [66]	0.8012	0.8360	0.8511	0.6613	0.8004	0.7490	0.6998	0.8666	0.8079	0.9023	0.9014	0.7933	0.7468	0.9627	0.8128
GLoRIA [25]	0.8263	0.8326	0.8596	**0.6641**	0.8179	0.7348	0.7104	0.8452	0.8129	0.8977	**0.9310**	0.7886	0.7608	0.9750	0.8184
BioViL [6]	0.8185	0.8543	0.8607	0.6660	0.8302	0.7633	0.7090	0.8595	**0.8287**	0.9031	0.9251	0.7912	0.7638	0.9696	0.8245
ours	**0.8291**	**0.8594**	**0.8719**	0.6565	**0.8382**	**0.7647**	**0.7378**	**0.8807**	0.8275	**0.9083**	0.9224	**0.7977**	**0.7784**	**0.9796**	**0.8323**

表12：在 ChestX-ray 14 种疾病分类任务上，与其他最先进方法进行微调比较。报告了每种疾病的 AUC 分数，同时也报告了宏平均 AUC 分数。我们用疾病名称的前三个字母表示疾病，但对“肺炎”和“气胸”则分别使用前两个字母和最后一个字母。

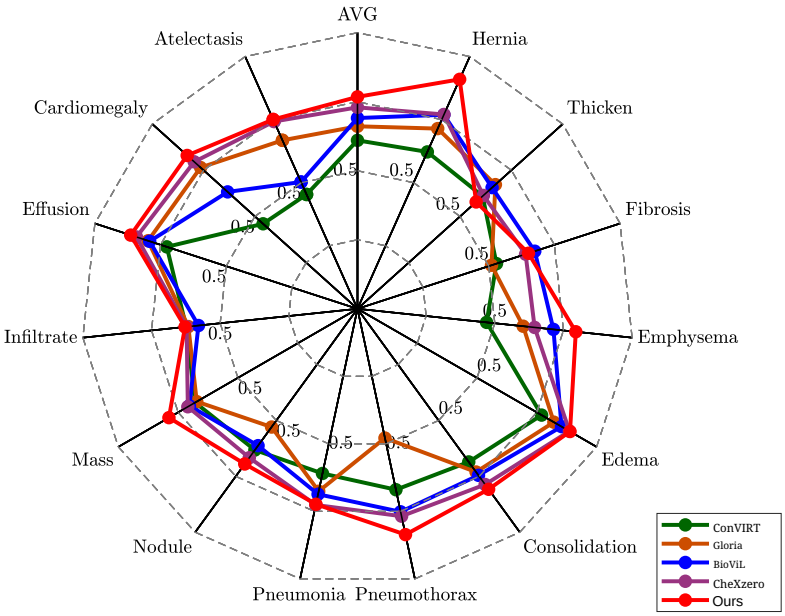


图3：我们的方法与其他方法在 ChestX-ray14 的 14 种疾病上的雷达图。图中报告了 AUC 分数，如图所示，我们的方法在大多数疾病上均优于此前的最先进方法。

E. 可视化结果

图 4 展示了我们的模型在零样本定位任务中的可视化结果。如图所示，“肺炎”的真实标注以边界框给出，通常对应较大区域。因此，该数据集上的指标高于另外两个数据集。我们的网络能够很好地捕捉其区域。对于“气胸”，其异常模式与其他疾病不同，目标是捕捉肺部萎陷部分，在图像上呈现较暗区域，而非较亮的阴影。其真实标注掩膜通常细长狭窄，但我们的网络仍能突出显示其位置。对于“Covid-19”，其图像文本与“肺炎”相似，但由于这是一种全新的疾病，定位其区域更具挑战性。这要求模型根据它们复杂的定义和症状建立相互关系。可视化结果表明，我们的模型成功实现了这一点，这也支持了我们的模型能够理解其他未见疾病的复杂描述。

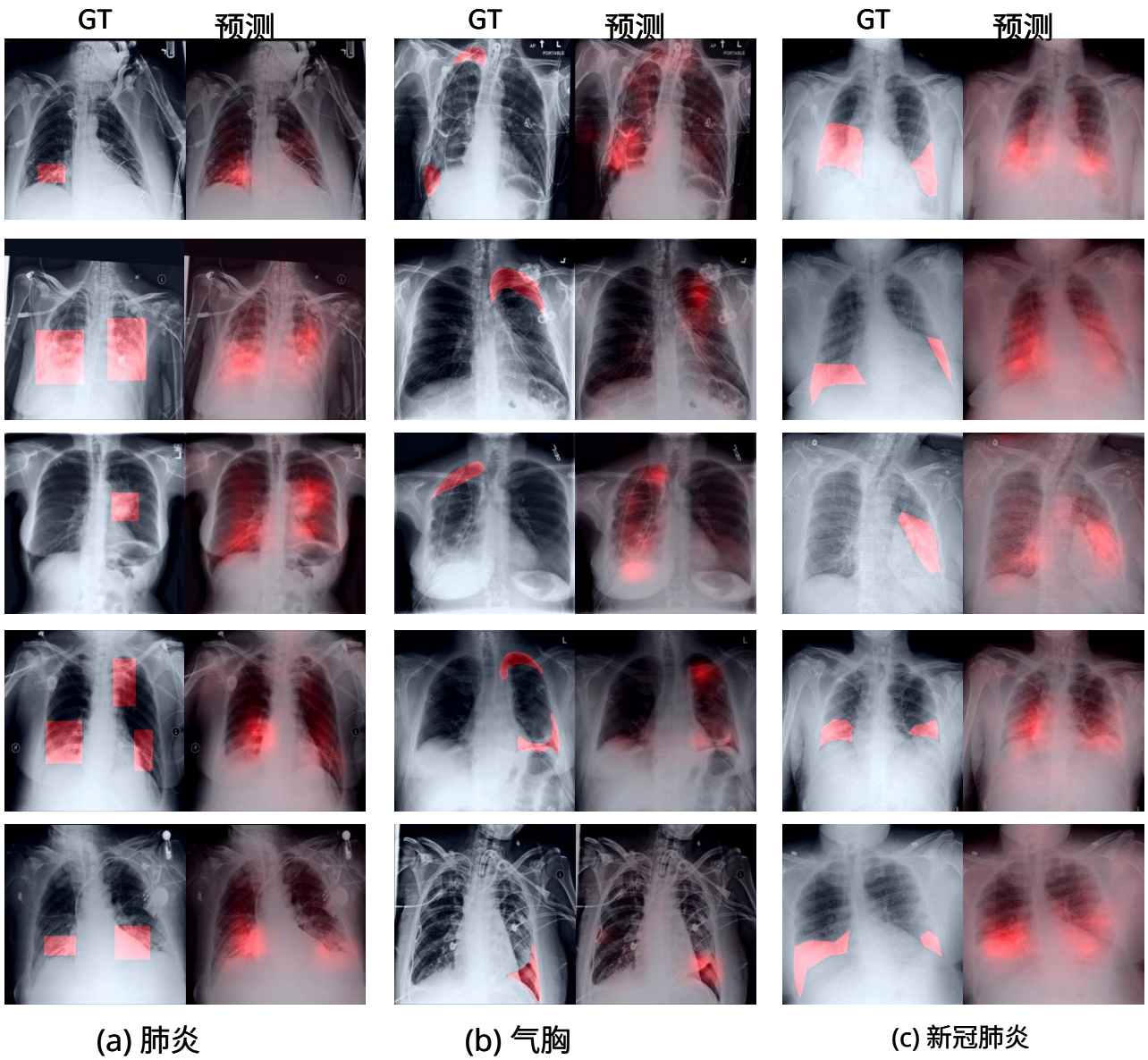


图 4：我们方法的零样本定位结果可视化。每列展示一种疾病的结果，左侧为真实标注，右侧为模型预测的热力图。图中红色越亮，表示模型越可能认为该区域与异常相关。