arXiv:2303.07240v1 [cs.CV] 13 Mar 2023

PMC-CLIP：使用生物医学文献进行对比式语言-图像预训练

WeixiongLin[1,*]，Ziheng Zhao[1,*]，Xiaoman Zhang[1,2]，Chaoyi Wu[1,2]，YaZhang[1,2]，Yanfeng Wang[1,2]，和 Weidi Xie[1,2,†]

[1] 上海交通大学协同媒体网创新中心，中国上海[2] 上海人工智能实验室，中国上海{wxlin,Zhao Ziheng,xm99sjtu,wtzxxxwcy02,ya zhang,wangyanfeng,weidi}@sjtu.edu.cn

摘要。在大规模数据集上训练的基础模型近期在计算机视觉和自然语言处理领域取得了迅速发展。相比之下，由于数据稀缺，生物医学领域的发展远远落后。为解决这一问题，我们构建并发布了 PMC-OA，这是一个生物医学数据集，包含从 PubMed Central 开放获取子集中收集的 160 万组图像-描述对，规模是此前数据集的 8 倍。PMC-OA 涵盖多种模态和疾病，其中大多数图像说明以更细粒度的层次对齐，*i.e.*、子图和子图注。在 PMC-OA 上预训练 CLIP 风格模型后，我们提出的 PMC-CLIP 在多种下游任务上取得了最先进的结果，包括 ROCO 上的图文检索、MedMNIST 图像分类、医学视觉问答、 *i.e.* +81% 的图文检索 R@10、 +3.9% 的图像分类准确率。

关键词： 基础模型 · 多模态数据集 · 视觉-语言预训练。

1 引言

近年来，基础模型的发展一直是人工智能的主要驱动力，例如，通过自回归预测或掩码词元填充进行 [26,8,2,23]训练的大语言模型，以及 [17,25,30] 通过对比视觉-语言特征进行训练的计算机视觉模型。相比之下，受两个方面的数据可用性限制，生物医学领域的发展远远落后：（i）标注所需的专业知识；（ii）隐私问题。本文介绍了我们的初步研究，旨在利用公开可用的科学论文构建一个**大规模、高质量的图文**生物医学数据集，且**仅需投入少量人工**工作。

具体而言，我们从 PubMed Central 上的科学文献中抓取图和对应的图像说明。PubMed Central 是美国国立卫生研究院旗下美国国家医学图书馆提供的生物医学与生命科学期刊文献免费全文档案，

†：通讯作者。*：这些作者对本研究的贡献相同。

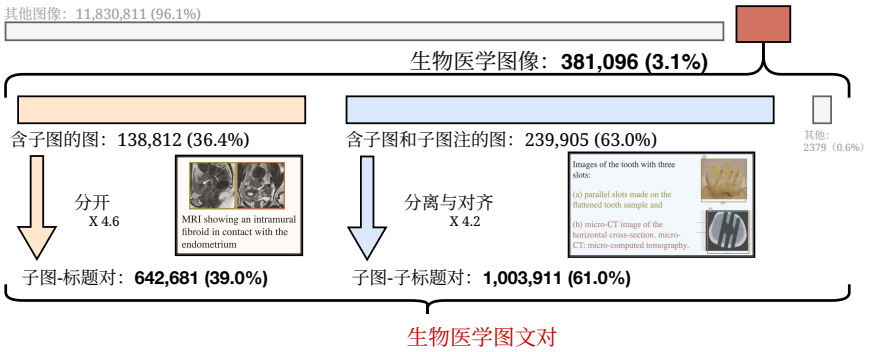


图 1. PMC-OA 收集流程及相关统计数据。

美国国家医学图书馆（美国国立卫生研究院/美国国家医学图书馆） [27]。这带来了两项优势：（i）出版物中的内容通常经过充分标注并由专家审核；（ii）图均已妥善匿名化并去除身份信息。在文献领域，我们显然并非首个以这种方式构建生物医学数据集的团队；然而，现有数据集相较于当今标准存在一些局限。例如，作为开创性工作，ROCO [24] 很久以前构建，仅包含 8.1 万张放射影像。MedICaT [29] 包含 21.7 万张图像，但大多由复合图构成。在本研究中，我们通过引入一套自动化流程来解决上述局限，该流程从科学文献中生成具有子图与子图注对应关系的数据集，包含三个主要阶段：医学图像收集、子图分离、子图注分离与对齐。最终数据集 PMC-OA 包含 165 万个图文对，见图1 和图3。

在构建数据集的同时，我们还训练了一个面向生物医学领域的 CLIP 式视觉语言模型，称为 PMC-CLIP。该模型在 PMC-OA 上使用标准的图文对比（ITC）损失进行训练，并且为了促进图像与文本的联合交互，还采用了掩码语言建模（MLM）。我们在多个下游任务上评估了预训练模型，包括医学图文检索、医学图像分类和医学视觉问答（VQA）。PMC-CLIP 在各种下游任务上均取得了最先进的性能，显著超越了以往的方法。

总而言之，本文的贡献如下：**首先**，我们提出了一套自动化流程，用于从科学论文中构建高质量的图文生物医学数据集，并通过该流程构建了一个名为 PMC-OA 的图像描述数据集，其规模比以往大 $8\times$ 。借助所提出的流程，该数据集可以持续更新。**其次**，我们在构建的图像描述数据集上预训练了一个名为 PMC-CLIP 的视觉语言模型，使其成为生物医学领域的基础模型。**第**三，我们在多种任务（检索、分类和视觉问答）上进行了全面实验，并在大多数下游数据集上取得了最先进性能，展现了 PMC-OA 的优越性以及基础模型 PMC-CLIP 的潜力。该数据集和预训练模型将向社区开放。

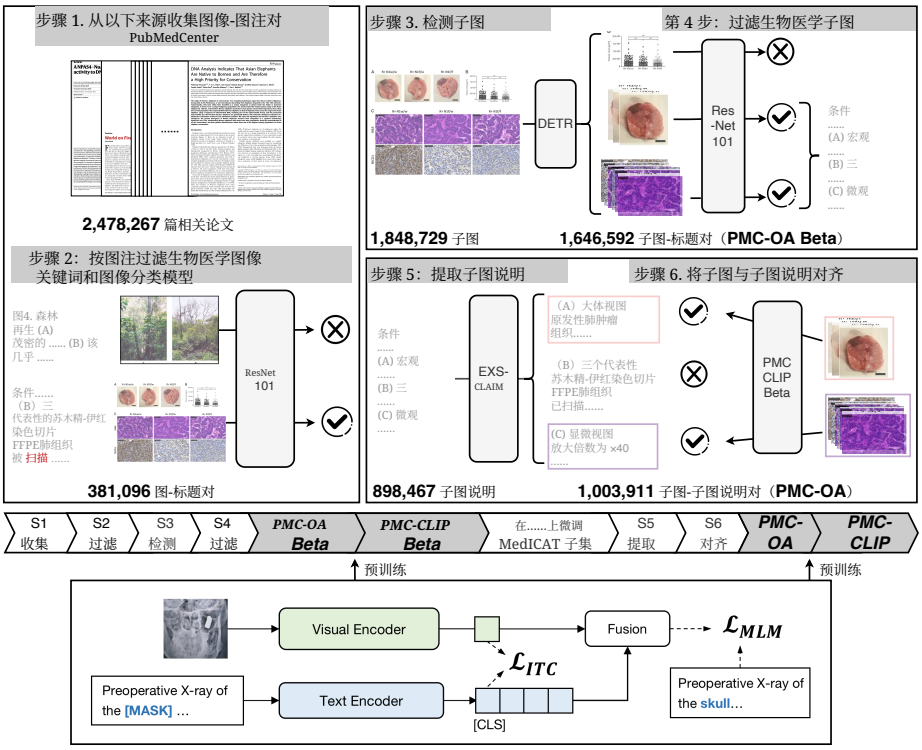


图 2. 用于收集 PMC-OA（上方）的流程，以及 PMC-CLIP（下方）的架构。

2 PMC-OA 数据集

本节首先介绍第2.1节中的数据集收集流程，随后简要概述第2.2节中的 PMC-OA。

2.1 数据集收集

本节详细介绍构建 PMC-OA 的流程。PMC-OA 是一个包含 165 万组图文对的大规模数据集。整个过程分为三个主要阶段：(i) 医学图像收集，(ii) 子图分离，(iii) 子图注分离与对齐，如图 2 所示。

医学图像收集（图 2中的步骤 1 和 2）。我们首先从 PubMed Central 提取图和图注。截至 2022-09-16，涵盖了 2,478,267 篇可用论文，并提取了 12,211,907 对图像与图注。为获得医学图像，我们遵循 [29]：首先，预先定义一组医学关键词 来过滤图注，删除

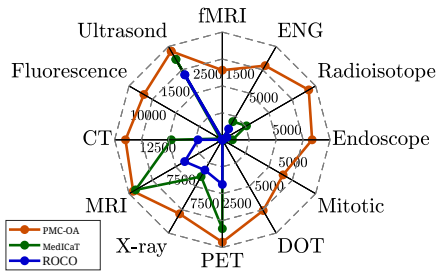
遵循 UMLS [1] 中定义的 T060 类“*Diagnostic Procedure*”

图注中未出现任何关键词的图文对；其次，使用在 DocFigure 上训练的 ResNet-101 [12] [13] 进行科学图表分类，将剩余的图分为 28 个类别。我们保留“医学”类别预测得分在 28 个类别中排名前 4 的图，最终得到 381,096 张医学图。

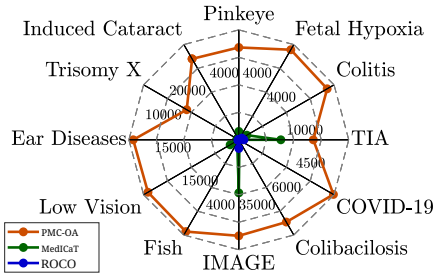
子图分离（图2中的步骤 3 和 4）。我们随机抽查了上一步得到的约 300 张图，发现其中约 80% 是复合图，即包含多个面板。在此，我们训练一个检测器，将复合图拆分为子图。具体而言，我们使用一个基于 ResNet-34 的 DETR 模型 [3]，该模型包含 4 个编码器层、4 个解码器层和 32 个可学习查询。该检测器在 MedICaT 子集 [29]，上进行训练，该子集包含经人工分离的 2,069 张复合图及其对应图注，以下称为 MedICaTSub。为进行超参数调优，我们按 3:1 的比例将数据集划分为训练集和测试集，模型在测试集上的 mAP@0.5 达到 0.94。为平衡精确率（0.93）和召回率（0.94），我们将置信度阈值设为 0.7。拆分复合图后，图表等非医学子图可能会与医学子图混在一起，因此我们使用分类模型对生成的子图进行过滤（重复第一阶段的操作）。至此，我们从 378,717 张复合图中获得了 1,646,592 张子图；图注数量与复合图相同，因此目前每条图注平均对应 4.3 张子图。我们将此数据集称为 **PMC-OA-贝塔** 版本。

子图注分离与对齐（图 2 中的步骤 5 和 6）。为了进一步将子图与完整图像说明中对应的部分（即子图注）对齐，我们首先需要将图像说明拆分为子图注。为此，我们使用 [28] 中提供的图注拆分工具处理 PMC-OA Beta 中的所有图像说明。**请注意**，该工具有时无法拆分图像说明，因此我们从 239,905 条可拆分的图像说明中获得了 898,467 条子图注，对应 1,003,911 个子图。为了将每个子图与最相关的子图注对齐，我们在 **PMC-OA-Beta** 上预训练了一个 CLIP 风格模型（训练详情将在第 3 节介绍），然后在 MedICaTSub 上对其进行微调。具体而言，对于每个子图，我们使用对比学习微调预训练模型，使其与对应的子图注匹配。我们按 3:1 的比例将该子集划分为训练集和测试集，微调后的模型在测试集上的对齐准确率达到=73%。最终，我们对齐了 1,003,911 对子图-子图说明对。将其与剩余的 642,681 对子图-图像说明对合并后，我们将该数据集命名为 **PMC-OA**。随后，我们在该数据集上预训练了 **PMC-CLIP** 。

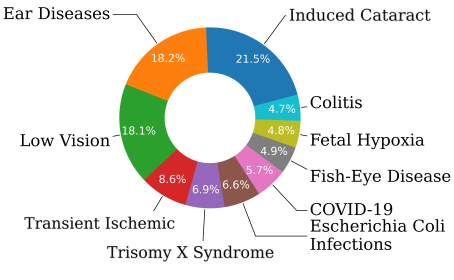
讨论。 一些开创性工作 [24,29]已探索公开可用的科学文献，以缓解生物医学领域的数据稀缺问题。尽管这些工作为社区提供了宝贵的数据集，但仍存在一些局限。我们的工作旨在改进数据收集流程，以构建规模更大、更多样且更准确的数据集： 首先，PMC-OA 涵盖的论文数量（2,478,267）多于 ROCO [24]（1,828,575）和 MedICaT [29]（131,410），因此数据集规模也更大（160 万）。其次，与 ROCO [24], 不同，我们保留了非放射学图像，因此 PMC-OA 包含更多样的生物医学数据。图 3 展示了定量比较。 第三，据我们所知，我们首次将子图分离、子标题分离及其对齐整合到数据收集流程中，从而显著扩大了数据集规模（约为 MedICaT 的 8 倍，ROCO 的 20 倍），同时尽可能减少噪声。



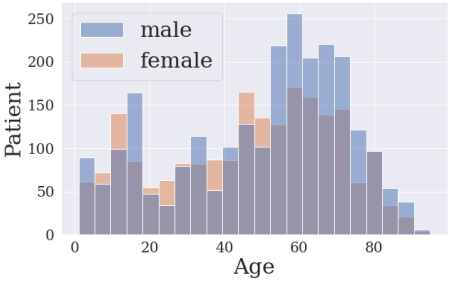
(a) 诊断操作。



(b) 疾病与检查结果。



(c) PMC-OA 中的疾病分布。



(d) PMC-OA 中患者的年龄与性别。

图 3. PMC-OA 的统计概览。

约为 MedICaT 的 8 倍和 ROCO 的 20 倍，同时尽可能减少噪声。

2.2 数据集概览

本节从三个不同方面对收集到的 PMC-OA 数据集进行简要统计概览，即诊断操作、疾病与检查结果，以及公平性。

诊断操作。如图所示，3(a)，PMC-OA 涵盖了范围广泛的诊断操作，从常见的（*CT*、*MRI*、*X*射线）到罕见的（有丝分裂象），比以往多样得多。

疾病与检查结果。在 PMC-OA 中，疾病以自由文本形式呈现，便于进行细致的识别和分析。例如，眼部疾病可以进一步分为白内障、结膜炎、黄斑变性、鱼眼病等。图3(b)展示了诊断中常用的词语，图3(c) 展示了它们的分布。

公平性。我们在图中展示了不同年龄段的性别比例。3(d)，可以看到，PMC-OA 的性别比例大致均衡，其中男性占 54%。

讨论。详细的数据集统计结果从三个方面体现了 PMC-OA 的优势：（i）诊断程序多样性；（ii）疾病覆盖范围；（iii）人群公平性。首先，迄今为止，应用最广泛的文本-图像数据集是 MIMIC-CXR[14]，它仅包含胸部 X 光图像，极大地限制了视觉语言预训练方法的潜力，而我们的数据集可以弥补这一不足。其次，诊断始终是临床中的关键程序，而我们的数据集广泛的疾病覆盖范围有助于学习疾病的共性模式，从而促进准确的自动诊断。第三，人群公平性确保我们的数据集受到的患者特征偏差影响较小，因此具有更强的跨中心泛化能力。

3 视觉语言预训练

利用我们构建的图像描述数据集，我们进一步训练了一个视觉语言模型，称为 PMC-CLIP，如图2 （下图）所示。我们在第3.1 节介绍其架构，并在第3.2节介绍训练目标

3.1 架构

给定一个包含 N 个图像-描述对的数据集，即即 $\mathcal{D} = \{(\mathcal{I}_1, \mathcal{T}_1) \cdots, (\mathcal{I}_N, \mathcal{T}_N)\}$，其中 $\mathcal{I}_i \in \mathbb{R}^{H\times W\times C}$ 表示图像， H, W, C 分别表示高度、宽度和通道， $\mathcal{T}_i$ 表示与图像配对的文本。我们的目标是训练一个 CLIP 风格视觉语言模型，其中包含图像编码器 Φ_{visual} 和文本编码器 Φ_{text}。

具体而言，给定一个图像-描述对（$\mathcal{I}, \mathcal{T}$），我们分别使用基于 ResNet 的 Φ_{visual} 和基于 BERT 的 Φ_{text} 对其进行编码；嵌入维度记为 d，文本词元长度记为 l:

$$\boldsymbol{v} = \Phi_{\text{visual}}(\mathcal{I}) \in \mathbb{R}^d, \tag{1}$$

$$\boldsymbol{T} = \Phi_{\text{text}}(\mathcal{T}) \in \mathbb{R}^{l\times d},\ \boldsymbol{t} = \boldsymbol{T}_0 \in \mathbb{R}^d, \tag{2}$$

其中， $\boldsymbol{v}$ 表示整幅图像的嵌入， $\boldsymbol{T}$ 表示句子嵌入，而 $\boldsymbol{t}$ 表示 [CLS] 标记的嵌入。

3.2 训练目标

在本节中，我们使用两个目标训练视觉语言模型，即图文对比学习和掩码语言建模。

图文对比学习（ITC）。我们按照 CLIP 的方法实现图文对比损失 [25], ，旨在匹配同一样本中对应的视觉和文本表征。具体而言，设批次大小为 b，我们计算当前视觉/文本嵌入（$\boldsymbol{v}$ / $\boldsymbol{t}$）与批次内所有样本之间经 softmax 归一化的跨模态点积相似度，分别记为 p^{i2t}， $p^{\text{t2i}} \in \mathbb{R}^b$，最终的图文对比损失为:

$$L_{\text{ITC}} = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{D}}\left[\text{CE}(y^{\text{i2t}}, p^{\text{i2t}}) + \text{CE}(y^{\text{t2i}}, p^{\text{t2i}})\right], \tag{3}$$

其中， y^{i2t}， y^{t2i} 表示独热匹配标签，CE 表示 InfoNCE 损失 [22]。

掩码语言建模（MLM）。我们按照 BERT [7] 实现 MLM 损失。网络经过训练，可根据上下文内容和视觉线索重建被掩码的词元。我们以 15% 的概率随机掩蔽文本中的词，并将其替换为词元 [MASK]。我们将图像嵌入 $\boldsymbol{v}$ 与句子嵌入 $\boldsymbol{T}$ 拼接后，输入自注意力 Transformer 融合模块 Φ_{fusion}，并在输出序列的相应位置获得被掩码词元的预测，记为 $p^{\mathrm{mask}} = \Phi_{\mathrm{fusion}}(\boldsymbol{v}, \boldsymbol{T})$。令 y^{mask} 表示真实值，则 MLM 损失为：

$$L_{\mathrm{MLM}} = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{D}}\left[\mathrm{CE}(y^{\mathrm{mask}}, p^{\mathrm{mask}})\right] \tag{4}$$

总训练损失。最终损失是上述两项损失的加权和：

$$L = L_{\mathrm{ITC}} + \lambda L_{\mathrm{MLM}}, \tag{5}$$

其中， λ 是决定 L_{MLM} 权重的超参数，默认设为 0.5。

4 实验设置

这里，我们首先介绍第4.1节中用于比较的数据集，接着在第4.2节中描述下游任务，并在第4.3节中介绍实现细节。

4.1 预训练数据集

ROCO [24]是一个从 PubMed 收集的图像-描述数据集 [27]。它过滤掉所有复合图像或非放射学图像，包含 81K 个样本。

MedICaT [29] 将 ROCO 扩展至 217K 个样本（图像-描述对），但其图中有 75% 是复合图，即 一张图包含多个子图。

MIMIC-CXR [14]是规模最大的胸部 X 射线数据集，包含 377,110 个样本（图像-报告对）。每张图像都配有一份由医生撰写、描述检查结果的临床报告。

4.2 下游任务

图文检索（ITR）。ITR 包含图像到文本（I2T）和文本到图像（T2I）检索。我们在不同的数据集上训练 PMC-CLIP，并在 ROCO 测试集上进行评估。**请注意**，我们已明确排除自身数据与 ROCO 之间的重复项，因此报告的结果相当于零样本评估。遵循以往研究 [29,4,5], ，我们从 ROCO 测试集中抽取 2,000 个图文对进行评估。

分类。 我们针对不同的下游任务对模型进行微调，这些任务侧重于图像分类。具体而言，MedMINIST [31] 共包含 12 项任务，涵盖生物医学图像中的主要数据模态，包括 ColonPathology、Dermatoscope、视网膜光学相干断层扫描等。

视觉问答（VQA）。 我们在 VQA-RAD [16], SLAKE [19], 的官方数据集划分上进行评估。SLAKE 由 642 张图像和 14,028 个问题组成，而 VQA-RAD 包含 315 张图像和 3,515 个问题。VQA-RAD 和 Slake 中的问题分为封闭式和开放式：若答案选项有限，则为封闭式；否则为开放式。在将 PMC-CLIP 适配到视觉问答任务时，我们保留了大部分预训练参数，包括视觉编码器、文本编码器和融合模块。具体而言，将图像和问题输入 PMC-CLIP，然后使用融合模块输出的嵌入，计算其与各个候选答案（也由文本编码器编码）之间的相似度。为取得更好的适配效果，我们分别在文本编码器中使用 10 个可学习的答案提示向量和问题提示向量。

指标。 我们使用准确率和曲线下面积评估分类任务，使用召回率@K（K=1、5、10）评估检索任务，并使用准确率评估视觉问答。

4.3 实现细节

对于视觉编码器和文本编码器，我们采用 ResNet50 [12] 和 PubmedBERT [11]。融合模块使用 4 个 Transformer 层。对于输入数据，我们将每张图像调整为 224×224。在预训练期间，文本编码器以 PubmedBERT 初始化，而视觉编码器和融合模块则从头开始训练。我们使用 AdamW [20] 优化器，参数为 $lr = 1 \times 10^{-4}$。我们在 GeForce RTX 3090 GPU 上训练，批次大小为 128，共训练 100 轮。前 10 轮用于预热。

5 结果

我们通过实验验证所提出数据集的有效性，以及在该数据集上训练的模型的效果。在第 5.1 节中，我们首先在图文检索任务上与现有的大规模生物医学数据集进行比较，以展示 PMC-OA 的优势。在第 5.2 节中，我们在三个不同的下游任务（检索、分类和视觉问答）上对在 PMC-OA 上预训练的模型进行微调。此外，我们还在第 5.3 节中，对预训练目标和模型架构进行了全面的实证研究。**请注意**，除非另有说明，所有实验均采用默认设置：图像编码器使用 ResNet50，并使用图文对比和掩码语言建模目标进行预训练。

5.1 PMC-OA 超越最先进水平的大规模生物医学数据集

如表 1所示，我们在不同数据集上对 PMC-CLIP 进行预训练，并在 ROCO 测试集上评估检索性能。仅切换为我们的数据集，性能便大幅提升，证实了该数据集的重要性。

表1. 关于预训练数据集的消融研究。

方法	预训练数据	数据规模	I2T			T2I		
			R@1	R@5	R@10	R@1	R@5	R@10
PMC-CLIP	ROCO	173 K	12.30	35.28	46.52	13.36	35.84	47.38
PMC-CLIP	MedICaT	17.3 万	12.30	35.28	46.52	13.36	35.84	47.38
PMC-CLIP	PMC-OA Beta	160 万	30.42	59.11	70.16	27.92	55.99	66.35
PMC-CLIP	PMC-OA	160 万	31.41	61.15	71.88	28.02	58.33	69.69

5.2 PMC-CLIP在下游任务中达到最先进水平

为了评估 PMC-CLIP 学到的表征，我们在多个下游任务中将其与若干最先进的方法进行比较，包括图文检索、图像分类和视觉问答。

图文检索。如表2所示，我们在图文检索方面取得了最先进的结果。在图像到文本排名@10指标上，PMC-CLIP的表现比此前的最先进方法高出8.1%。值得一提的是，M3AE [4], ARL [5]的预训练使用了ROCO的训练集，而我们的数据集不包含来自ROCO的数据。

表2。ROCO上的零样本图文检索。深灰色和浅灰色分别突出显示各项指标中的最佳和次佳结果。

方法	预训练数据	数据规模	I2T			T2I		
			R@1	R@5	R@10	R@1	R@5	R@10
ViLT [15]	COCO、VG、SBU、GCC	4.1M	11.90	31.90	43.20	9.75	28.95	41.40
METER [9]	COCO、VG、SBU、GCC	4.1M	14.45	33.30	45.10	11.30	27.25	39.60
M3AE [4]	ROCO、MedICaT	233 K	19.10	45.60	61.20	19.05	47.75	61.35
ARL [5]	ROCO、MedICaT、CXR	233 K	23.45	50.60	62.05	23.50	49.05	63.00
PMC-CLIP	PMC-OA	1.6 M	31.41	61.15	71.88	28.02	58.33	69.69

图像分类。 为展示 PMC-CLIP 出色的迁移能力，我们在 MedMNIST 上对其进行微调，并与最先进水平的方法进行比较，即、DWT-CV [6] 和 SADAE [10]。此处展示 12 项子测试中的 3 项结果，完整结果见补充材料。如表3所示，PMC-CLIP 始终取得更高的结果。值得注意的是，与使用 ResNet 从头训练相比，从 PMC-CLIP 开始微调能显著提升性能。

表 3. MedMNIST 上的分类结果。深灰色和浅灰色分别突出显示各项指标中的最佳和次佳结果。

数据集	PneumoniaMNIST		BreastMNIST		DermaMNIST	
方法	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
ResNet50 [12]	96.20	88.40	91.90	86.60	91.20	73.10
DWT-CV [6]	95.69	88.67	89.77	85.68	91.67	74.75
SADAE [10]	98.30	91.80	91.50	87.80	92.70	75.90
PMC-CLIP	99.02	95.35	94.56	91.35	93.41	79.80

视觉问答。 视觉问答要求模型学习更细粒度的视觉和语言表征。如表4 所示，我们在 6 项结果中的 5 项上优于最先进水平的方法 M3AE。

表4。 VQA-RAD 和 Slake 上的视觉问答结果

方法	VQA-RAD			Slake		
	Open	已关闭	总体	Open	闭合	总体
MEVF-BAN [21]	49.20	77.20	66.10	77.80	79.80	78.60
CPRD-BAN [18]	52.50	77.90	67.80	79.50	83.40	81.10
M3AE [4]	67.23	83.46	77.01	80.31	87.82	83.25
PMC-CLIP	67.00	84.00	77.60	81.90	88.00	84.30

5.3 消融研究

为说明预训练目标（图文对比、掩码语言建模）的有效性，我们对在 PMC-OA 上预训练的 PMC-CLIP 进行了消融研究。为验证融合模块的必要性，我们比较了仅使用纯文本上下文的掩码语言建模（掩码语言建模-*T*）和结合视觉线索的掩码语言建模（视觉掩码语言建模）。

如表 5所示，我们可以得出以下观察结果：首先，图文对比目标对预训练至关重要，对性能的贡献最大（ID 1）。其次，仅使用文本上下文的掩码语言建模可作为正则化项（ID 2）。第三，加入视觉特征后，模型能够学习图像-描述对之间更细粒度的关联，并取得最佳结果（ID 3）。

表 5。 预训练目标的消融研究

ID	方法			ROCO				PneumoniaMNIST	BreastMNIST		DermaMNIST	
	图文对比	纯文本掩码语言建模	视觉掩码语言建模	R@1	R@5	R@10	曲线下面积	ACC	曲线下面积	准确率	曲线下面积	准确率
1	✓			25.53	53.48	64.64	98.07	93.56	93.04	89.74	92.77	77.91
2	✓	✓		29.17				94.31	94.39	90.71	92.76	78.00
3	✓		✓	29.72				95.35	94.56	91.35	93.41	79.80

6 结论

本文通过从大量科学文献中收集图像-描述对，构建了一个生物医学领域的大规模数据集，名为 PMC-OA。我们在 PMC-OA 上训练了一个 CLIP 风格模型，称为 PMC-CLIP。该模型在多种下游生物医学任务中达到了最先进水平，包括图像-文本检索、图像分类和视觉问答。借助自动化的数据收集流程，我们相信该数据集将有益于研究群体，并推动生物医学领域基础模型的发展。

参考文献

1. Olivier Bodenreider。统一医学语言系统（UMLS）：整合生物医学术语。*Nucleic acids research* 32 1 267 270, 2004 3，（增刊）：D–D。。

2 Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Pra-fulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等，语言模型是少样本学习者。*Advances in neural information processing systems* 33 1877 1901, 2020 ：–。1 3. Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用变换器进行端到端目标检测。载于 *Computer Vision–ECCV 2020: 16th European Conference*、*Glasgow UK August*、、、、*23–28 2020 Part I 16* 213 229 2020 *Proceedings* 4，第–页。施普林格，。。

4 Zhihong Chen、Yuhao Du、Jinpeng Hu、Yang Liu、Guanbin Li、Xiang Wan 和 Tsung-Hui Chang。用于医学视觉与语言预训练的多模态掩码自编码器。载于 *International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference Singapore* 、*September* *18–22 2022 Part V* 679 689 2022 、*Proceedings* 7, 9, 10，第–页。施普林格，。。

5 Zhihong Chen、Guanbin Li 和 Xiang Wan。对齐、推理与学习：利用知识增强医学视觉与语言预训练。载于 *30th ACM International Conference on Multimedia* 5152 5161, 2022 *Proceedings of the*，第–页。7, 9 6. Jianhong Cheng、Hulin Kuang、Qichang Zhao、Yahui Wang、Lei Xu、Jin Liu 和 Jianxin Wang。DWT-CV：一种基于密集权重迁移的交叉验证策略，用于生物医学数据分析中的模型选择。*Future Generation Computer Systems*，135 20 29，2022：–。9 7. Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。*preprint arXiv:1810.04805* 2018 *arXiv*。7 8. Ming Ding、Wendi Zheng、Wenyi Hong 和 Jie Tang。CogView2：通过分层变换器实现更快、更优的文本到图像生成。*arXiv:2204.14217* 2022 *arXiv preprint*。1 9. Zi-Yi Dou、Yichong Xu、Zhe Gan、Jianfeng Wang、Shuohang Wang、Lijuan Wang、Chenguang Zhu、Pengchuan Zhang、Lu Yuan、Nanyun Peng 等。端到端视觉与语言变换器训练的实证研究。载于 *Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第18166 9 10 18176页，2022–。。Xiaolong Ge、Yanpeng Qu、Changjing Shang、Longzhi Yang 和 Qiang Shen。一种用于医学应用的自适应判别自编码器。*IEEE Transactions on Circuits and Systems for Video Technology* 32 12 8875 8886, 2022 ，（）：–。9 11. Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。用于生物医学自然语言处理的领域特定语言模型预训练。*ACM Transactions on Computing for Healthcare (HEALTH)* 3 1 1 23, 2021 ，（）：–。8 12. Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于 *Proceedings of the IEEE conference on computer vision and pattern recognition* 770 778, 2016 ，第–页。4, 8, 9 13. KV Jobin、Ajoy Mondal 和 CV Jawahar。DocFigure：科学文献图表分类数据集。载于 *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* 1, 74 79 ，卷，第–页。IEEE, 2019。4

14. Alistair EW Johnson、Tom J Pollard、Seth J Berkowitz、Nathaniel R Greenbaum、Matthew P Lungren、Chih-ying Deng、Roger G Mark 和 Steven Horng。Mimic-cxr：一个去标识化、公开可用的胸部X线片及自由文本报告数据库。 *Scientific data* 6 1 317, 2019 6, 7 , ( ): . . Wonjae Kim 、 Bokyung Son 和 Ildoo Kim。ViLT：无需卷积或区域监督的视觉与语言变换器。载于 *Learning* 5583 5594 2021 , 第 – 页。 15 *International Conference on Machine* PMLR, . 9 16. Jason J Lau、Soumya Gayen、Asma Ben Abacha 和 Dina Demner-Fushman。一个关于放射影像的临床生成视觉问答数据集。 *Scientific data* 5 1 1 10, 2018 8 , ( ): – . . Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。先对齐后融合：通过动量蒸馏进行视觉与语言表征学习。 17 *Advances in neural information processing systems* 34 9694 9705, 2021 , : – . 1 18. Bo Liu、Li-Ming Zhan 和 Xiao-Ming Wu。基于放射影像的医学视觉问答对比预训练与表征蒸馏。载于 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021:* , , , , , *24th International Conference Strasbourg France September 27–October 1 2021 Proceedings Part II 24* 210 220 2021 , , 第 – 页。施普林格，. 10 19. Bo Liu、Li-Ming Zhan、Li Xu、Lin Ma、Yan Yang 和 Xiao-Ming Wu。Slake：用于医学视觉问答的语义标注知识增强数据集。载于 *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1650 1654 2021 , 第 – 页。IEEE，. 8 20. Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。 *arXiv preprint arXiv:1711.05101* 2017 , . 8 21. Binh D Nguyen、Thanh-Toan Do、Binh X Nguyen、Tuong Do、Erman Tjiputra 和 Quang D Tran。克服医学视觉问答中的数据受限问题。载于 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019:* , , , , , *22nd International Conference Shenzhen China October 13–17 2019 Proceedings Part IV 22* 522 530 2019 , , 第 – 页。施普林格，. 10 22. Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。使用对比预测编码进行表征学习。 *arXiv preprint arXiv:1807.03748*, 2018 7 23 . . Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。训练语言模型，使其通过人类反馈遵循指令。 *arXiv preprint arXiv:2203.02155* 2022 , . 1 24. Obioma Pelka、Sven Koitka、Johannes R¨uckert、Felix Nensa 和 Christoph M Friedrich。语境中的放射学对象（ROCO）：一个多模态图像数据集。载于 *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018 Granada Spain September 16 2018 Proceedings* , , , , , , 第 – 页。施普林格，. 2, 4, 7 25. Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sand-hini 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 *3* 180 189 2018 *International conference on machine learning* 8748 8763 2021 , 第 – 页。PMLR，. 1, 6 26. Aditya Ramesh 、 Prafulla Dhariwal 、 Alex Nichol 、 Casey Chu 和 Mark Chen。使用 CLIP 潜变量进行分层文本条件图像生成。 *arXiv preprint arXiv:2204.06125* 2022 , . 1 27. Richard J Roberts。PubMed Central：已发表文献的 GenBank，2001。2, 7

28. Eric Schwenker、Weixin Jiang、Trevor Spreadbury、Nicola Ferrier、Oliver Cossairt 和 Maria KY Chan。Exsclaim!——一个从文献构建标注材料成像数据集的自动化流程。 *arXiv preprint arXiv:2103.10631*, 20214。29. Sanjay Subramanian、Lucy Lu Wang、Sachin Mehta、Ben Bogin、Madeleine van Zuylen、Sravanthi Parasa、Sameer Singh、Matt Gardner 和 Hannaneh Hajishirzi。MedICaT：一个包含医学图像、图像说明和文本引用的数据集。*preprint arXiv:2010.06000* 2020 arXiv ，。

2，3，4，7 30. Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhiliang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mohammed、Saksham Singhal、Subhojit Som 等。图像即外语：面向所有视觉及视觉语言任务的 Beit 预训练。 *arXiv preprint arXiv:2208.10442*, 20221 31 . . Jiancheng Yang、Rui Shi、Donglai Wei、Zequan Liu、Lin Zhao、Bilian Ke、Hanspeter Pfister 和 Bingbing Ni。Medmnist v2——用于二维和三维生物医学图像分类的大规模轻量级基准。 *Scientific Data*, 10(1):41, 2023。7