

基于预训练方法的视觉-语言模型 (VLMs)

Patrick Sylvestre

Department of Automation, Tsinghua University

November 25, 2024

Abstract

在经典的基于深度神经网络 (DNNs) 的视觉任务中，无论其范式如何，往往都是一个数据密集型的任务。往往完成一个较好的视觉任务模型都需要对每一个任务进行单独的数据收集和训练，显然这需要消耗大量的人力物力，并且在某些场景下由于数据的缺失，这个任务无法被很好完成。为解决这两个问题，视觉-语言模型 (Vision-Language Models, VLMs) 通过互联网上几乎无限的图像文本对数据中学习丰富的视觉语言相关性，并且展现了极强的零样本泛化能力，因此近年来引起了着重的关注。VLMs 的训练方式主要分为三种：基于预训练、基于迁移学习和基于知识蒸馏，由于预训练方案目前具有最好的效果，因此以下讨论的均是围绕基于预训练的 VLMs 展开的。本文参考综述 [67, 111, 1, 92]，结合最新至 2024 年 11 月的工作，将从 VLMs 的发展背景、基本原理与训练方法、训练/评估数据集、模型性能以及未来发展路径几个方面叙述探讨 VLMs 技术。

1 什么是 VLMs

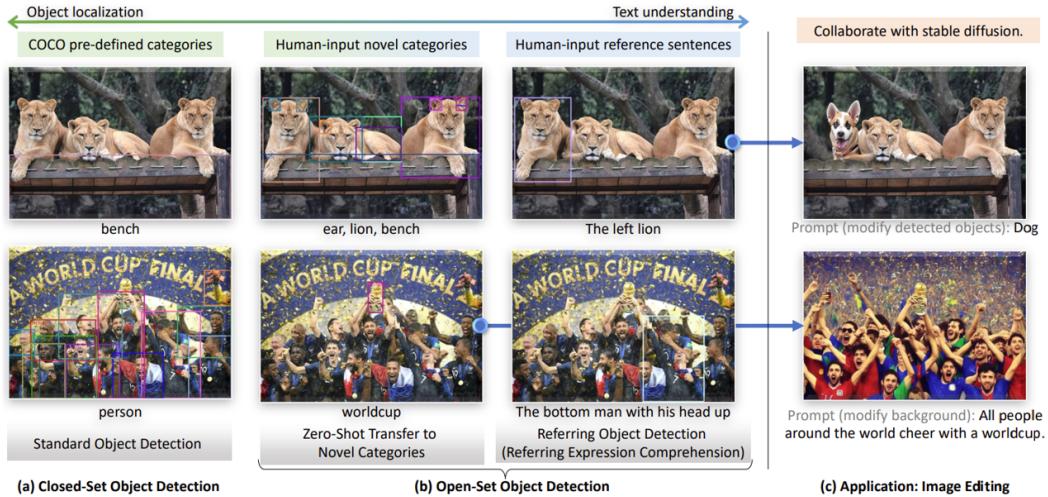


Figure 1: 图 (a) 表示传统范式的闭集物体检测需要模型检测预定义类别的物体。图 (b) 表示在未定义的新对象上使用 VLMs 模型的预测结果，该评估标准指称表达理解 (Referring Expression Comprehension, REC) 基准，以了解模型对具有属性的新型物体的泛化能力。图 (c) 是结合 VLMs 和 Stable Diffusion[81] 提出了一种图像编辑方案，也即是“可理解即可编辑”。图引自 [61]。

视觉识别（主要包括但不限于：图像分类、目标检测和语义分割等任务）是计算机视觉研究中的一个长期挑战，同时也是众多计算机视觉应用（如自动驾驶、遥感、医疗图像、工

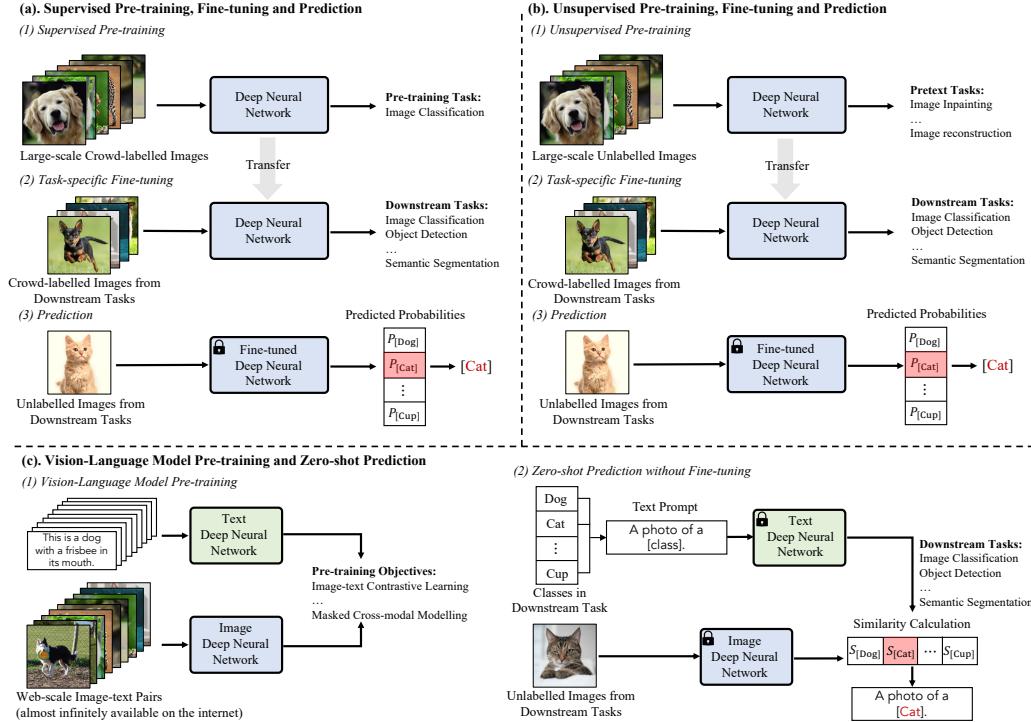


Figure 2: 三种视觉识别中的 DNN 训练范式。与需要对每个具体任务进行带任务特定标注数据微调的 (a) 和 (b) 两种范式相比，基于 VLMs 的新学习范式 (c) 实现了网络数据的高效利用和无需任务特定微调的零样本预测。图引自 [111]。

业质检、具身智能等) 的基石。随着深度学习的兴起，通过端到端可训练的深度神经网络 (DNNs)，视觉识别研究取得了巨大成功。然而，从传统机器学习向深度学习的转变也带来了两个新的重大挑战：即从头开始训练 (training from scratch) 的经典设置下 DNN 训练收敛缓慢；以及在 DNN 训练中需要耗费大量人力收集大规模的特定任务数据，并且需要使用例如众包等方式进行手工标注 [76]。

稍早几年，一种新的学习范式预训练、微调和预测在多种视觉识别任务中展示了极大的有效性 [35, 41, 12]。在这一新范式下，DNN 模型首先使用某些现成的大规模训练数据 (标注或未标注) 进行预训练，然后使用特定任务的标注训练数据对预训练模型进行微调，如图 2(a) 和 (b) 所示。通过在预训练模型中学习到的全面知识，这一学习范式可以加速网络收敛，并为各种下游任务训练表现优异的模型。

然而尽管如此，预训练、微调和预测范式仍然需要一个额外阶段，对每个下游任务使用带标注的训练数据进行任务特定的微调。受到自然语言处理进展的启发 [24, 77, 78]，一种新的深度学习范式——视觉-语言模型预训练与零样本预测最近引起了越来越多的关注 [76, 46, 106] (如图 1 所示)。在这一范式中，VLMs 通过互联网上几乎无限可用的大规模图文对进行预训练，预训练的 VLMs 可以直接应用于下游视觉识别任务，而无需微调，如图 2(c) 所示。VLMs 预训练通常通过某些视觉-语言目标 [76, 106, 108] 指导，从大规模图文对中学习图文对应关系 [82, 83]，例如，CLIP [76] 采用图文对比目标，通过在 embedding¹ 空间中拉近配对图文的距离并推远其他图文的距离进行学习。通过这种方式，预训练的 VLMs 捕捉到了丰富的视觉-语言对应知识，可以通过匹配任意给定图像和文本的 embedding 进行零样本预测。这种新学习范式能够高效利用网络数据，并无需任务特定微调即可实现零样本预测，简单易实现且性能出色，例如，预训练的 CLIP 在 36 项视觉识别任务中 (从经典的图像分类 [4, 54, 98, 52, 73] 到人体动作和光学字符识别 [50, 76, 66, 90, 8]) 实现了卓越的零样本性能。

在视觉-语言模型预训练与零样本预测取得巨大成功之后，研究人员对两条研究方向展开了深入探索，超越了各种 VLMs 预训练的研究范围。第一条方向探索了通过迁移学习 (Transfer Learning) 来使用 LMs [118, 117, 33, 112]。许多迁移方法已被提出，如提示调优 (Prompt

¹有时可以译为“特征”，但不贴切。这里如同 Token 一样，不好翻译为中文。

Tuning) [118, 117]、视觉适配 (Visual Adaptation) [33, 112] 等，它们的共同目标是将预训练的 VLMs 高效地适配到各种下游任务。第二条方向则探索了基于知识蒸馏 (knowledge Distilling) 的 VLMs 研究 [25, 37, 28]，例如，一些研究 [25, 37, 28] 致力于从 VLMs 中提取知识应用于下游任务，以期在目标检测、语义分割等任务中获得更好的性能。但是篇幅所限，这两个方向不在本文的讨论范围内。

尽管近年来与 VLMs 相关的研究论文数量激增，整个领域处于急速发展的状态，所有研究者都在为这个“长久以来的梦想”而努力。但该技术对传统的技术路线造成了巨大的冲击，以至于此前大量基于预训练、微调和预测范式的方法有可能会直接过时。本文的内容参考综述 [67, 111, 1, 92]，结合最近最近的工作整理而来。主要对 VLMs 在视觉识别任务中的应用（包括图像分类、目标检测和语义分割）和截至 2024 年 11 月的最新进展（主要是基于预训练方法）进行了系统性的整理和讨论，并探讨了 VLMs 面临的若干研究挑战及困难。

2 VLMs 的发展背景

本节首先介绍视觉识别训练范式的发展历程，并描述其如何演进为视觉-语言模型预训练与零样本预测范式。随后，我们介绍视觉-语言模型 (VLMs) 在视觉识别中的发展。

2.1 视觉识别的训练范式

视觉识别范式的发展可大致分为五个阶段，包括：

1. 传统机器学习范式
2. 经典深度学习范式
3. 有监督预训练、微调范式
4. 无监督预训练、微调范式
5. 视觉-语言模型预训练与零样本预测

以下内容将对这五种训练范式进行简单的介绍、比较和分析。

传统机器学习范式由来已久，但主要在 2014 年在深度学习兴起之前作为主流方案，这种范式下视觉识别研究主要依赖于特征工程，即手工设计特征 [65, 62]，以及 SVM、KNN、随机森林等轻量机器学习模型 [5, 74, 19]，将手工设计的特征分类为预定义的语义类别。好处是可解释性强，模型性能有一定程度的数学保证。然而，这种范式需要领域专家为特定的视觉识别任务设计有效的特征，不仅难以应对复杂任务，而且扩展性较差。

经典深度学习范式是随着深度学习的兴起发展而来，这种范式下视觉识别研究通过使用端到端 (end-to-end) 可训练的深度神经网络 (DNNs) 取得了巨大成功。这种方法避免了复杂的特征工程，将研究转向神经网络架构的架构工程以学习有效的特征。例如，ResNet[42] 通过残差结构使网络可以更深，从而能够利用大规模的标注数据进行学习，在具有挑战性的 ImageNet 基准测试中表现出色 [21]。然而，从传统机器学习向深度学习的转变引发了两个新的巨大挑战：经典深度学习训练收敛速度较慢，以及大规模、任务特定、人工标注数据的收集成本高昂 [76]。

为克服上述经典深度学习范式的缺点，研究发现，从大规模标注数据中学习的特征可以迁移到下游任务 [35]，**有监督预训练、微调范式**新范式逐渐取代了此前的范式，如图 2 (a) 所示，通过使用监督损失在大规模标注数据集（例如 ImageNet）上预训练 DNNs，并在任务特定的训练数据上微调预训练的 DNN[35]。由于预训练的 DNN 已经学习了某些视觉知识，可以加速网络的收敛，并帮助在有限的任务特定训练数据下训练出性能优异的模型。

无监督预训练、微调范式 [41, 12] 的新学习范式的提出，探索使用自监督学习从未标注数据中学习有用且可迁移的表示，如图 2 (b) 所示，尝试解决了预训练阶段需要大规模标注数据的问题。基于此范式，提出了多种自监督训练目标 [40, 41]，包括建模跨区域关系的掩码图像建模 [40]、通过对比训练样本学习判别特征的对比学习 [41] 等。这些自监督预训练的模型随后通过任务特定的标注数据进行下游任务微调。由于此范式在预训练阶段无需标注数据，可以利用更多的数据进行训练，从而学习到更有用且可迁移的特征，表现甚至优于有监督预训练 [12, 41]。

但无论怎样，预训练与微调范式（有监督或无监督预训练）仍然需要利用带标注的任务数据进行微调，本质上仍没有解决特定数据来源的问题。受自然语言处理领域巨大成功的启

发 [24, 77, 78]，一种全新的深度学习范式被提出，即视觉-语言模型预训练与零样本预测范式，如图 2 (c) 所示。借助互联网上几乎无限可用的大规模图像-文本对，VLMs 通过特定的视觉-语言目标 [76, 106, 108] 进行预训练，从而捕获丰富的视觉-语言知识，并能够通过匹配任意给定图像和文本的 embedding，在下游视觉识别任务中实现零样本预测（无需微调）。这一特性使得视觉任务在使用端或许真的能够摆脱数据密集型的要求，蜕变为真正的“智能”模型。

与预训练与微调范式相比，这一新范式能有效利用大规模网络数据并在无需任务特定微调的情况下进行零样本预测。现有研究主要从三个方面改进 VLMs：

- 收集大规模的有信息图像-文本数据；
- 设计高容量模型以有效从大数据中学习；
- 设计新的预训练目标以学习有效的 VLMs。

2.2 VLMs 在视觉领域的发展

自 OpenAI 的开创性工作 CLIP [76] 的发表以来，视觉识别相关的 VLMs 研究取得了显著进展。如图 3 所示，以下三个方面展示了视觉识别中 VLMs 的发展：

1. **预训练目标**从“单一目标”到“多目标混合”：早期的 VLMs（例如 [76, 46]）通常采用单一的预训练目标，而近期的 VLMs（例如 [89, 106]）引入了多种目标（例如，对比、对齐和生成目标），探索这些目标之间的协同作用以构建更稳健的 VLMs 并提升下游任务性能。
2. **预训练框架**从“多独立网络”到“统一网络”：早期的 VLMs（例如 [76, 46]）采用双塔式的预训练框架，而近期的 VLMs（例如 [95, 45]）尝试单塔式的预训练框架，该框架通过统一网络对图像和文本进行编码，不仅减少了 GPU 内存使用，还提高了跨数据模态的通信效率。
3. **下游任务**从简单任务到复杂任务：早期的 VLMs（例如 [76, 46]）专注于图像级视觉识别任务，而近期的 VLMs（例如 [105, 63]）则更为通用，不仅可以处理图像级任务，还能够支持需要定位相关知识的密集预测任务，这类任务更加复杂。

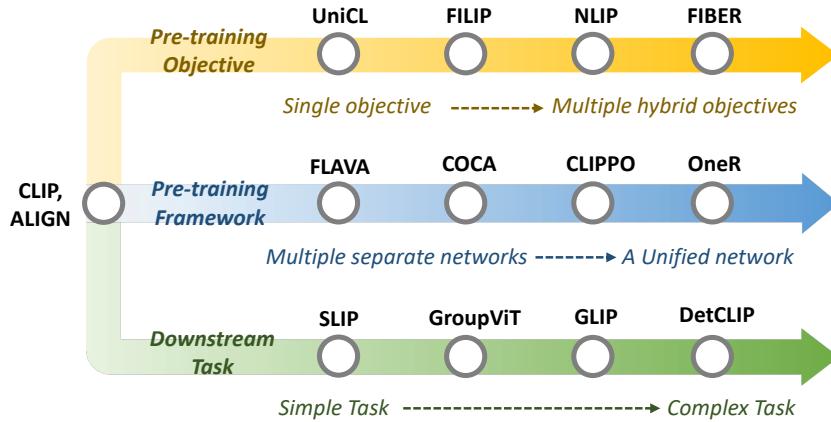


Figure 3: VLMs 在视觉任务上的发展路线。图引自 [111]。

3 VLMs 的基本原理与训练方法

VLMs 预训练 [76, 46] 旨在预训练 VLMs，以学习图像-文本关联，从而在视觉识别任务中实现高效的零样本预测。在给定图像-文本对 [83, 82] 的情况下，首先使用文本编码器 [96, 24] 和图像编码器 [42, 26] 提取图像和文本特征，然后通过特定的预训练目标学习视觉-语言关联 [76, 46]。因此，VLMs 可以通过匹配任意给定图像和文本的 embedding，在零样本的情况下评估未见数据的性能 [76, 46]。本节介绍了 VLMs 预训练的基础知识，包括用于提取图像和文本特征的常见网络架构、建模视觉-语言关联的预训练目标、VLMs 预训练框架以及用于评估 VLMs 的下游任务。

3.1 网络架构

VLMs 预训练依赖深度神经网络从预训练数据集 $\mathcal{D} = \{x_n^I, x_n^T\}_{n=1}^N$ 中的 N 个图像-文本对中提取图像和文本特征，其中 x_n^I 和 x_n^T 分别表示一个图像样本及其配对的文本样本。深度神经网络包含一个图像编码器 f_θ 和一个文本编码器 f_ϕ ，它们将图像-文本对 x_n^I, x_n^T 编码为图像 embedding $z_n^I = f_\theta(x_n^I)$ 和文本 embedding $z_n^T = f_\phi(x_n^T)$ 。本节展示了 VLMs 预训练中广泛采用的深度神经网络架构。

3.1.1 用于学习图像特征的架构

用于学习图像特征的网络架构主要分为两类：基于卷积神经网络（CNN）的架构和基于 Transformer 的架构。

基于 CNN 的架构：不同的卷积网络（例如，VGG[88]、ResNet[42] 和 EfficientNet[93]）被设计用于学习图像特征。ResNet[42] 是 VLMs 预训练中最受欢迎的卷积网络之一，通过在卷积块之间引入 Skip Connections，大幅缓解了梯度消失和爆炸问题，使得训练非常深的神经网络成为可能。为了改进特征提取和视觉-语言建模，一些研究 [76] 对原始网络架构 [42, 93] 进行了修改。例如，针对 ResNet，研究者引入了 ResNet-D[43]，在 [113] 中使用了抗锯齿 rect-2 模糊池化，并将全局平均池化替换为 Transformer 多头注意力 [96] 中的注意力池化。

基于 Transformer 的架构：近年来，Transformer 在视觉识别任务中被广泛研究，包括图像分类 [26]、目标检测 [7] 和语义分割 [100] 等。作为用于图像特征学习的标准 Transformer 架构，ViT[26] 由一系列 Transformer 块组成，每个块包含一个多头自注意力层和一个前馈网络（FFN）。输入图像首先被分割成固定大小的图块，然后在经过线性投影和位置 embedding 后输入 Transformer 编码器。相关研究（例如 [76, 69, 106]）通过在 Transformer 编码器前添加归一化层对 ViT 进行改进。

3.1.2 用于学习语言特征的架构

Transformer 及其变体 [96, 78, 24] 广泛用于学习文本特征。标准 Transformer[96] 具有一个编码器-解码器结构，其中编码器包含 6 个块，每个块由一个多头自注意力层和一个多层次感知机（MLP）组成；解码器也包含 6 个块，每个块包含一个多头注意力层、一个掩码多头注意力层和一个 MLP。大多数 VLMs 研究（例如 CLIP[76]）采用了标准 Transformer[96]，并在 GPT2[78] 的基础上进行轻微修改，直接从头开始训练，而不依赖 GPT2 的预训练权重初始化。

3.2 VLMs 的预训练目标

作为 VLMs 的核心，多种视觉-语言预训练目标 [76, 41, 103, 24, 39, 108, 89, 58] 被设计用于学习丰富的视觉-语言关联。这些目标大致分为三类：对比目标（Contrastive Objectives）、生成目标（Generative Objectives）和对齐目标（Alignment Objectives）。

3.2.1 对比目标

对比目标通过在特征空间中拉近配对样本、推远非配对样本 [76, 41, 103]，训练 VLMs 学习判别性特征。

图像对比学习（Image Contrastive Learning）：图像对比学习旨在通过将 query 图像与其正样本（即数据增强的图像）拉近、与负样本（即其他图像）推远，来学习判别性图像特征 [41, 12]。给定一个包含 B 张图像的批次，对比学习目标（例如 InfoNCE[71] 及其变体 [12, 41]）通常定义如下：

$$\mathcal{L}_I^{\text{InfoNCE}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^I \cdot z_+^I / \tau)}{\sum_{j=1, j \neq i}^{B+1} \exp(z_i^I \cdot z_j^I / \tau)}, \quad (1)$$

其中， z_i^I 是 query embedding， $\{z_j^I\}_{j=1, j \neq i}^{B+1}$ 是 key embedding， z_+^I 表示 z_i^I 的正样本 key，其余为负样本 key。 τ 是控制特征密度的超参数。

图像-文本对比学习（Image-Text Contrastive Learning）：图像-文本对比学习旨在通过拉近成对图像与文本 embedding、推远非成对 embedding[76, 46]，学习判别性图像-文本特

征。这通常通过最小化对称的信息最大化 (infoNCE) 损失 $\mathcal{L}_{\text{infoNCE}}^{IT} = \mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}$ 来实现，其中 $\mathcal{L}_{I \rightarrow T}$ 和 $\mathcal{L}_{T \rightarrow I}$ 分别对比图像与文本和文本与图像。给定包含 B 个图像-文本对的批次， $\mathcal{L}_{I \rightarrow T}$ 和 $\mathcal{L}_{T \rightarrow I}$ 定义如下：

$$\mathcal{L}_{I \rightarrow T} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^I \cdot z_i^T / \tau)}{\sum_{j=1}^B \exp(z_i^I \cdot z_j^T / \tau)}, \quad (2)$$

$$\mathcal{L}_{T \rightarrow I} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^T \cdot z_i^I / \tau)}{\sum_{j=1}^B \exp(z_i^T \cdot z_j^I / \tau)} \quad (3)$$

图像-文本-标签对比学习 (Image-Text-Label Contrastive Learning): 图像-文本-标签对比学习 [103] 将监督对比学习 [49] 引入图像-文本对比学习，通过调整式 (2) 和 (3) 如下定义：

$$\mathcal{L}_{I \rightarrow T}^{ITL} = -\sum_{i=1}^B \frac{1}{|\mathcal{P}(i)|} \sum_{k \in \mathcal{P}(i)} \log \frac{\exp(z_i^I \cdot z_k^T / \tau)}{\sum_{j=1}^B \exp(z_i^I \cdot z_j^T / \tau)}, \quad (4)$$

$$\mathcal{L}_{T \rightarrow I}^{ITL} = -\sum_{i=1}^B \frac{1}{|\mathcal{P}(i)|} \sum_{k \in \mathcal{P}(i)} \log \frac{\exp(z_i^T \cdot z_k^I / \tau)}{\sum_{j=1}^B \exp(z_i^T \cdot z_j^I / \tau)}, \quad (5)$$

其中 $k \in \mathcal{P}(i) = \{k | k \in B, y_k = y_i\}$ [103]， y 是 (z^I, z^T) 的类别标签。最终损失定义为 $\mathcal{L}_{\text{infoNCE}}^{ITL} = \mathcal{L}_{I \rightarrow T}^{ITL} + \mathcal{L}_{T \rightarrow I}^{ITL}$ 。

3.2.2 生成目标

生成目标通过训练网络生成图像/文本数据（例如，图像生成 [41, 2]、语言生成 [24, 108] 或跨模态生成 [89]）来学习语义特征。

掩码图像建模 (Masked Image Modelling): 掩码图像建模通过随机掩码图像的一部分补丁并利用未掩码的补丁进行重建 [40, 2]，学习跨补丁的关联。给定一个包含 B 张图像的批次，损失函数定义如下：

$$\mathcal{L}_{MIM} = -\frac{1}{B} \sum_{i=1}^B \log f_\theta(\bar{x}_i^I | \hat{x}_i^I), \quad (6)$$

其中 \bar{x}_i^I 和 \hat{x}_i^I 分别表示图像 x_i^I 中被掩盖的补丁和未被掩盖的补丁。

掩码语言建模 (Masked Language Modelling): 掩码语言建模是自然语言处理 (NLP) 中广泛采用的预训练目标 [24]。它随机掩盖输入文本中一定比例（例如，BERT[24] 中为 15%）的文本 tokens²，并通过未掩盖的 tokens 进行重建：

$$\mathcal{L}_{MLM} = -\frac{1}{B} \sum_{i=1}^B \log f_\phi(\bar{x}_i^T | \hat{x}_i^T), \quad (7)$$

其中 \bar{x}_i^T 和 \hat{x}_i^T 分别表示文本 x_i^T 中被掩盖和未被掩盖的 tokens， B 表示批次大小。

掩码跨模态建模 (Masked Cross-Modal Modelling): 掩码跨模态建模结合了掩码图像建模和掩码语言建模 [89]。给定一个图像-文本对，它随机掩盖部分图像补丁和文本 tokens，并在未掩盖的图像补丁和文本 tokens 的条件下对其进行重建：

$$\mathcal{L}_{MCM} = -\frac{1}{B} \sum_{i=1}^B [\log f_\theta(\bar{x}_i^I | \hat{x}_i^I, \hat{x}_i^T) + \log f_\phi(\bar{x}_i^T | \hat{x}_i^I, \hat{x}_i^T)], \quad (8)$$

其中 \bar{x}_i^I/\hat{x}_i^I 分别表示图像 x_i^I 的掩盖/未掩盖补丁， \bar{x}_i^T/\hat{x}_i^T 分别表示文本 x_i^T 的掩盖/未掩盖 tokens。

²同 embedding 一样，不好翻译为中文。“Tokens are not real things, they’re a computer generated illusion created by gifted engineers.”

图像到文本生成 (Image-to-Text Generation): 图像到文本生成旨在基于与 x^T 配对的图像自回归地预测文本 x^T [108]:

$$\mathcal{L}_{ITG} = - \sum_{l=1}^L \log f_\theta(x^T \mid x_{<l}^T, z^I), \quad (9)$$

其中 L 表示需要预测的 x^T 的 tokens 数量, z^I 是与 x^T 配对的图像的 embedding。

3.2.3 对齐目标

对齐目标通过在 embedding 空间上进行全局图像-文本匹配或局部区域-单词匹配, 来对齐图像-文本对 [27, 3, 58, 105]。

图像-文本匹配 (Image-Text Matching) 图像-文本匹配建模图像和文本之间的全局相关性 [27, 3], 其公式基于评分函数 $\mathcal{S}(\cdot)$ 来度量图像和文本的对齐概率, 并结合二分类损失定义如下:

$$\mathcal{L}_{IT} = p \log \mathcal{S}(z^I, z^T) + (1-p) \log(1 - \mathcal{S}(z^I, z^T)), \quad (10)$$

其中, 当图像和文本配对时, $p = 1$, 否则 $p = 0$ 。

区域-单词匹配 (Region-Word Matching): 区域-单词匹配建模图像-文本对中“图像区域”和“单词”之间的局部跨模态相关性 [58, 105], 用于目标检测等密集视觉识别任务。公式定义如下:

$$\mathcal{L}_{RW} = p \log \mathcal{S}^r(r^I, w^T) + (1-p) \log(1 - \mathcal{S}^r(r^I, w^T)), \quad (11)$$

其中 (r^I, w^T) 表示一个区域-单词对, 若区域和单词配对, 则 $p = 1$, 否则 $p = 0$ 。 $\mathcal{S}^r(\cdot)$ 是用于度量“图像区域”和“单词”之间相似度的局部评分函数。

3.3 VLMs 的预训练框架

本节介绍了 VLMs 预训练中广泛采用的框架, 包括双塔、双腿和单塔预训练框架。

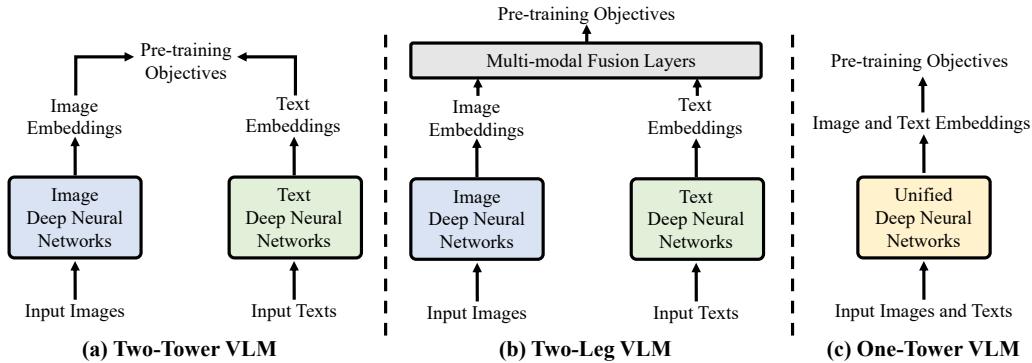


Figure 4: 典型 VLMs 预训练框架的示意图。图引自 [111]。

具体来说, 双塔框架在 VLMs 预训练中被广泛采用 [76, 46], 其中输入的图像和文本分别由两个独立的编码器进行编码, 如图 4 (a) 所示。略有不同的是, 双腿框架 [89, 108] 引入了额外的多模态融合层, 从而实现图像和文本模态特征的交互, 如图 4 (b) 所示。相比之下, 单塔 VLMs [95, 45] 尝试在一个单一的编码器中统一视觉和语言学习, 如图 4 (c) 所示, 其目标是促进数据模态间的高效通信。

3.4 评估设置与下游任务

本节介绍了 VLMs 评估中广泛采用的设置和下游任务。评估设置包括 **零样本预测 (Zero-shot Prediction)** 和 **线性探测 (Linear Probing)**, 下游任务则包括图像分类、目标检测、语义分割、图像-文本检索和动作识别等。

Table 1: VLMs 预训练中常用的图像-文本数据集总结。表中 [link] 可跳转至数据集官网。

数据集	年份	图像-文本对数量	语言	公开性
SBU Caption[72] [link]	2011	1M	英语	✓
COCO Caption[14] [link]	2016	1.5M	英语	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M)[94] [link]	2016	100M	英语	✓
Visual Genome (VG)[53] [link]	2017	5.4 M	英语	✓
Conceptual Captions (CC3M)[85] [link]	2018	3.3M	英语	✓
Localized Narratives (LN)[75] [link]	2020	0.87M	英语	✓
Conceptual 12M (CC12M)[10] [link]	2021	12M	英语	✓
Wikipedia-based Image Text (WIT)[91] [link]	2021	37.6M	108 种语言	✓
Red Caps (RC)[23] [link]	2021	12M	英语	✓
LAION400M[83] [link]	2021	400M	英语	✓
LAION5B[82] [link]	2022	5B	超过 100 种语言	✓
WuKong[36] [link]	2022	100M	中文	✓
CLIP[76]	2021	400M	英语	✗
ALIGN[46]	2021	1.8B	英语	✗
FILIP[106]	2021	300M	英语	✗
WebLI[13]	2022	12B	109 种语言	✗

3.4.1 零样本预测 (Zero-shot Prediction)

零样本预测是评估 VLMs 泛化能力的最常见方法 [76, 46, 106, 56, 69]，直接将预训练的 VLMs 应用于下游任务，而无需任何任务特定的微调 [76]。

图像分类 (Image Classification) [42, 88] 旨在将图像分类到预定义的类别中。VLMs 通过比较图像和文本的 embedding 来实现零样本图像分类，其中常采用“提示工程 (Prompt Engineering)”生成与任务相关的提示，例如“a photo of a [label].” [76]。

语义分割 (Semantic Segmentation) [11] 旨在为图像中的每个像素分配一个类别标签。预训练的 VLMs 可通过比较图像像素和文本的 embedding，实现分割任务的零样本预测。

目标检测 (Object Detection) [35, 80] 旨在定位和分类图像中的对象，这在各种视觉应用中十分重要。预训练的 VLMs 可通过比较目标提议 (object proposals) 和文本的 embedding，结合辅助数据集的学习能力 [84, 48]，实现目标检测的零样本预测。

图像-文本检索 (Image-Text Retrieval) [6] 旨在基于一种模态的线索从另一种模态中检索所需样本，包括文本检索图像和图像检索文本两项任务。

3.4.2 线性探测 (Linear Probing)

线性探测在 VLMs 评估中被广泛采用 [76]。其通过冻结预训练的 VLMs，仅训练一个线性分类器，将 VLMs 编码的 embedding 用于分类，以评估 VLMs 的表征能力。在图像分类 [42, 88] 和动作识别 [90, 66] 任务中，线性探测被广泛采用，其中动作识别任务通常对视频片段进行子采样以提高识别效率 [76]。

4 数据集

本节概述了 VLMs 预训练和评估中常用的数据集，详见表 1-2。

Table 2: VLMs 评估中广泛使用的视觉识别数据集概览。表中 [link] 链接到数据集官网。

任务	数据集	年份	类别数	训练集大小	测试集大小	评价指标
图像分类	MINIST[5] [link]	1998	10	60,000	10,000	Accuracy
	Caltech-101[32] [link]	2004	102	3,060	6,085	Mean Per Class
	Oxford 102 Flowers[70] [link]	2008	102	2,040	6,149	Mean Per Class
	CIFAR-10[54] [link]	2009	10	50,000	10,000	Accuracy
	CIFAR-100[54] [link]	2009	100	50,000	10,000	Accuracy
	ImageNet-1k[2] [link]	2009	1000	1,281,167	50,000	Accuracy
	Stanford Cars[52] [link]	2013	196	8,144	8,041	Accuracy
	FGVC Aircraft[64] [link]	2013	100	6,667	3,333	Mean Per Class
图像-文本检索	Food-101[4] [link]	2014	102	75,750	25,250	Accuracy
	Flickr30k[107] [link]	2014	-	31,783	-	Recall
动作识别	COCO Caption[14] [link]	2015	-	82,783	5,000	Recall
	UCF101[90] [link]	2012	101	9,537	1,794	Accuracy
	Kinetics[700] [link]	2019	700	494,801	31,669	Mean(top1, top5)
	RareAct[66] [link]	2020	122	7,607	-	mWAP, mSAP
目标检测	COCO 2014 Detection[60] [link]	2014	80	83,000	41,000	box mAP
	COCO 2017 Detection[60] [link]	2017	80	118,000	5,000	box mAP
	LVIS[38] [link]	2019	1203	118,000	5,000	box mAP
	ODInW[57] [link]	2022	314	132413	20070	box mAP
语义分割	PASCAL VOC 2012 Segmentation[29] [link]	2012	20	1464	1449	mIoU
	PASCAL Content[68] [link]	2014	459	4998	5105	mIoU
	Cityscapes[18] [link]	2016	19	2975	500	mIoU
	ADE20k[115] [link]	2017	150	25574	2000	mIoU

4.1 VLMs 的预训练数据集

用于 VLMs 预训练的多种大规模图像-文本数据集 [76, 46, 83, 82] 均来源于互联网。与传统人工标注数据集 [21, 18, 29] 相比，图像-文本数据集 [83, 76] 规模更大且成本更低。例如，最近的图像-文本数据集一般达到十亿级别规模 [83, 82, 13]。

除了图像-文本数据集，若干研究 [58, 108, 105, 95] 还利用辅助数据集提供额外信息以提升视觉语言建模效果，如 GLIP[58] 使用 Object365[84] 提取区域级特征。

4.2 VLMs 的测评数据集

如表 2 所示，许多数据集被用于评估 VLMs，包括 9 个最常见的图像分类数据集、4 个目标检测数据集、4 个语义分割数据集、2 个图文检索数据集以及 3 个动作识别数据集。例如，这 27 个图像分类数据集覆盖了广泛的视觉识别任务，从细粒度任务（如用于宠物识别的 Oxford-IIIT PETS 数据集 [73] 和用于汽车识别的 Stanford Cars 数据集 [52]）到一般任务（如 ImageNet 数据集 [21]）。

Table 3: VLMs 预训练方法总结。Con: 对比目标; Gen: 生成目标; Align: 对齐目标。†, ‡ 和 § 分别表示双塔、双腿和单塔预训练框架。* 表示非公开数据集。[\[code\]](#) 指向代码网站。

方法	数据集	目标	方法核心要点
CLIP†[76] [code]	CLIP*	Con	提出了一种基于对比学习的视觉语言预训练方法，利用图像和文本对进行语义匹配，大幅提升 VLMs 在多模态任务中的表现。
ALIGN†[46]	ALIGN*	Con	通过利用大规模的噪声图像-文本对数据，展示了数据规模在 VLMs 性能提升中的重要性。
OTTER†[97] [code]	CC3M, YFCC15M, WIT	Con, Gen	通过最优化传输方法提升数据效率，实现了更高效的 VLMs 预训练。
DeCLIP†[59] [code]	CC3M, CC12M, YFCC100M, WIT*	Con	在模型中融入图像和文本的自监督学习，提出了一种高效的数据使用策略。
ZeroVL†[20] [code]	SBU, VG, CC3M, CC12M	Con	引入数据增强技术，进一步提高了 VLMs 的训练效率。
FILIP†[106]	FILIP*, CC3M, CC12M, YFCC100M	Con, Align	通过引入区域-词汇相似性建模，实现了更精细的视觉语言预训练。
UniCL†[103] [code]	CC3M, CC12M, YFCC100M	Con	提出了一种结合图像-文本和标签的对比学习方法，统一了多模态任务的目标。
Florence†[109]	FLD-900M*	Con	扩展了预训练数据规模，并在模型中加入了深度和时间信息，增强了模型的多模态能力。
SLIP†[69] [code]	YFCC100M	Con	将图像的自监督学习方法引入 VLMs，进一步提升了模型表现。
PyramidCLIP†[34]	SBU, CC3M, CC12M, YFCC100M, LAION400M	Con	提出了跨层次对比学习方法，实现了更高的语义一致性。
ChineseCLIP†[102] [code]	LAION5B, WuKong, VG, COCO	Con	构建了大规模中文图像-文本数据集，并引入了针对中文的 VLMs。
LiT†[110] [project]	CC12M, YFCC100M, WIT*	Con	提出了锁定图像编码器的对比学习调优方法，提高了预训练效率。
AltCLIP†[115] [code]	WuDao, LAION2B, LAION5B	Con	使用多语言文本编码器，开发了支持多语言的 VLMs。
FLAVAV†[89] [code]	COCO, SBU, LN, CC3M, VG, WIT, CC12M, RC, YFCC100M	ALL	提出了通用基础模型 FLAVA，同时处理单模态和多模态任务。
KELIP†[51] [code]	CUB200, WIT, YFCC15M, CC3M, LAION400M, K-WIT*	Con, Gen	收集了大规模韩文图像-文本数据集，开发了支持韩语和英语的双语模型。
COCAT†[108] [code]	ALIGN*	Con, Gen	提出了结合对比学习和图像描述生成的预训练方法。
nCLIP†[116]	COCO, VG, SBU, CC3M, CC12M, YFCC14M	Con, Align	提出了非对比式预训练目标，改进了图像-文本匹配任务中的性能。
K-lite†[86] [code]	CC3M, CC12M, YFCC100M	Con	利用辅助数据集训练可迁移的 VLMs。
NLIP†[44]	YFCC100M, COCO	Con, Gen	通过降噪机制训练更加鲁棒的 VLMs。

5 VLMs 的预训练具体方法

在3.2小节中我们阐述了VLMs的预训练有三种典型训练目标：对比目标、生成目标和对齐目标。本节结合多项研究（详参表3）对这些方法更进一步的叙述。

5.1 基于对比目标的预训练

对比学习在VLMs预训练中被广泛研究，旨在通过设计对比目标学习辨别性的图像-文本特征[69, 59, 76]。

5.1.1 图像对比学习

该预训练目标旨在学习图像模态中的辨别性特征，通常作为辅助目标以充分挖掘图像数据潜力。例如，SLIP[69]使用公式1定义的标准infoNCE损失来学习辨别性的图像特征。

5.1.2 图像-文本对比学习

图像-文本对比学习旨在通过对比图像-文本对来学习视觉-语言相关性，即，通过拉近配对图像和文本的embedding并拉远其他非配对embedding的方式实现[76]。例如，CLIP[76]采用公式2中的对称图像-文本infoNCE损失来衡量图像与文本embedding的相似性（如图5）。该预训练模型因此能够学习图像-文本的相关性，从而在下游视觉识别任务中实现零样本预测。

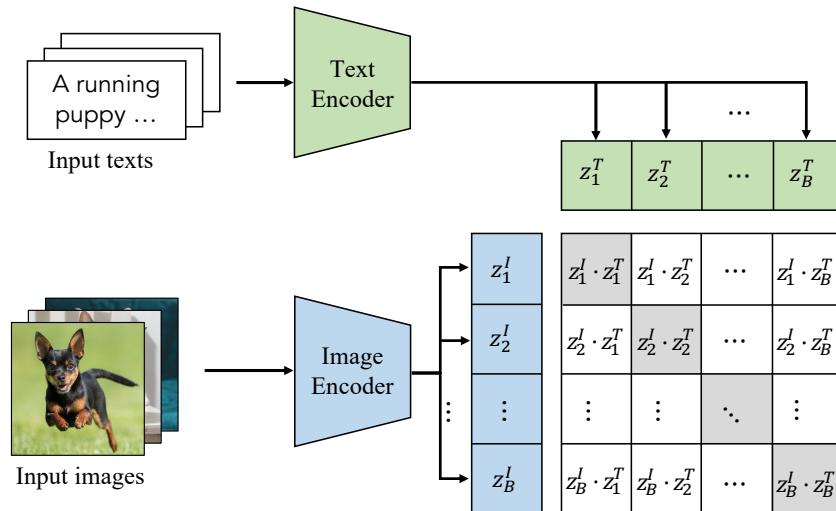


Figure 5: CLIP 中图像-文本对比学习的示意图。图引自[76]。

受CLIP成功的启发，许多研究从不同角度改进了对称图像-文本infoNCE损失。例如，ALIGN[46]在大规模（约18亿对）但噪声较大的图像-文本对中，采用抗噪对比学习来扩展VLMs预训练。一些研究[59, 97, 20]则探索了使用更少图像-文本对的数据高效预训练。例如，DeCLIP[59]引入最近邻监督，利用相似对的信息，在有限数据下实现有效的预训练；OTTER[97]使用最优传输技术伪配对图像和文本，大幅减少训练数据需求；ZeroVL[20]通过去偏的数据采样和coin flipping mixup的数据增强，充分利用有限的数据资源。

另一类研究[106, 34, 101]致力于通过在不同语义层面上执行图像-文本对比学习，实现全面的视觉-语言相关性建模。例如，FILIP[106]在对比学习中引入区域-单词对齐，从而学习细粒度的视觉-语言对应知识；PyramidCLIP[34]构建了多种语义层次，并在跨层和同层中执行对比学习以实现有效的预训练。

此外，近期的一些研究通过增强图像-文本对进一步改进[30, 99, 104, 22]。例如，LA-CLIP[30]和ALIP[104]利用大型语言模型为给定图像生成合成描述，而RA-CLIP[99]则检索相关的图像-文本对以增强数据。为了在数据模态间实现高效通信，[45]和[95]尝试在单一编码器中统一视觉和语言学习。

5.1.3 图像-文本-标签对比学习

该预训练方法在图像-文本对比中引入了图像分类标签 [103]，如公式 4 所定义的，目标是将图像、文本和分类标签编码到一个共享空间中（如图 6）。它结合了基于图像标签的有监督预训练和基于图像-文本对的无监督 VLMs 预训练。据 UniCL[103] 报告，该预训练方法能够同时学习辨别性和任务特定（如图像分类）的特征。随后在 [109] 的研究中，UniCL 扩展到约 9 亿对图像-文本对，从而在各种下游任务中表现出色。

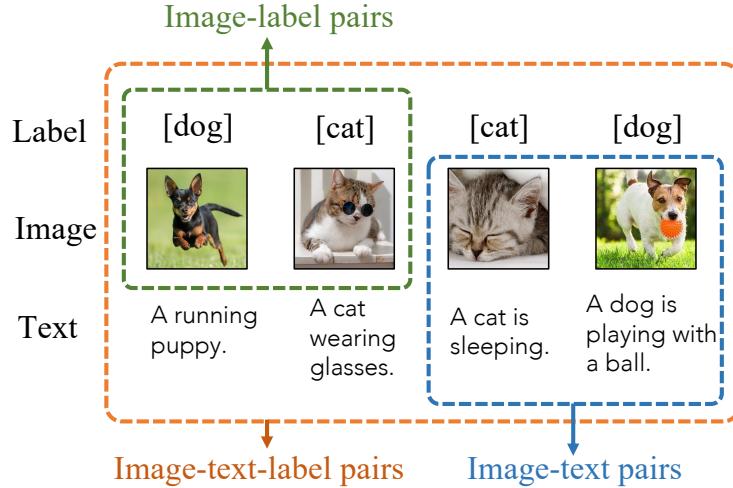


Figure 6: UniCL 提出的图像-文本-标签空间示意图。图引自 [103]。

5.1.4 讨论

对比目标通过拉近正样本对并远离负样本对的方式，促使模型学习辨别性的视觉和语言特征 [76, 46]，这种辨别性通常能提高模型零样本预测的置信度和准确性。然而，对比目标也存在两方面的局限性：联合优化正负样本对较为复杂且困难 [76, 46]；以及其引入了一个启发式温度超参数，用于控制特征的辨别性（详见3.2.1小节）。

5.2 基于生成目标的 VLMs 预训练

生成式 VLMs 预训练通过生成图像或文本来学习语义知识，主要采用的方式包括掩码图像建模、掩码语言建模、掩码跨模态建模和图像到文本生成。

5.2.1 掩码图像建模

该预训练目标通过掩码和重建图像来学习图像的上下文信息，定义如公式 6 所示。在掩码图像建模（如 MAE[40] 和 BeiT[2]）中，图像中的部分块被掩盖，编码器通过未掩盖的块进行重建学习，如图 7 所示。例如，FLAVA[89] 采用了 BeiT[2] 中的矩形块掩码方式，而 KELIP[51] 和 SegCLIP[63] 则借鉴了 MAE，训练中掩盖了大部分图像块（约 75%）。

5.2.2 掩码语言建模

掩码语言建模是 NLP 中广泛采用的预训练目标，其定义如公式 7 所示，也被证明在 VLMs 预训练中对文本特征学习十分有效。通过掩盖部分文本中的 tokens，并训练网络预测被掩盖的 tokens，如图 8 所示。例如，FLAVA[89] 参照 [24] 的做法，掩盖 15% 的文本 tokens，并通过其余 tokens 进行重建，以建模单词间的相关性。此外，FIBER[27] 也将掩码语言建模 [24] 作为 VLMs 预训练目标之一，用于提取更好的语言特征。

5.2.3 掩码跨模态建模

掩码跨模态建模通过同时掩盖和重建图像块和文本 tokens 来进行训练，其定义如公式 8 所示，继承了掩码图像建模和掩码语言建模的优点。该方法通过掩盖一定比例的图像块和文本 tokens，并利用未掩盖的图像块和文本 tokens 的 embedding 进行重建。例如，FLAVA[89]

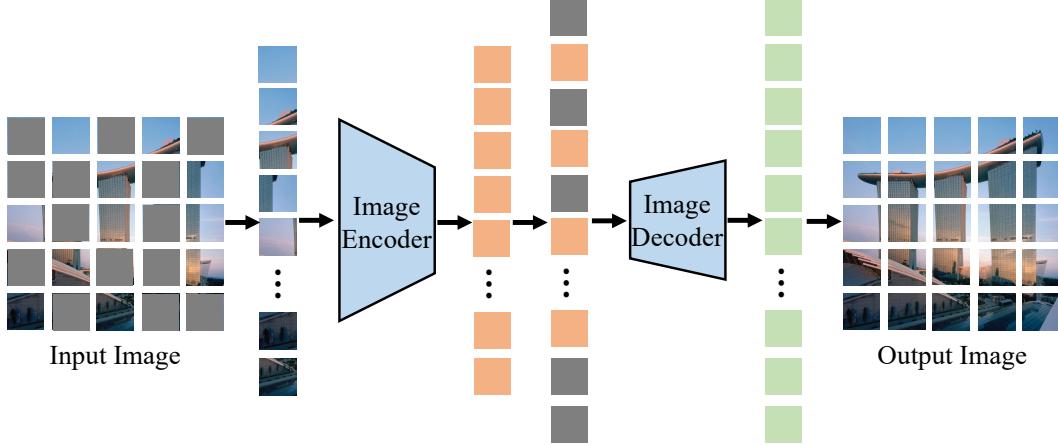


Figure 7: 掩码图像建模的示意图。图引自 [39]。

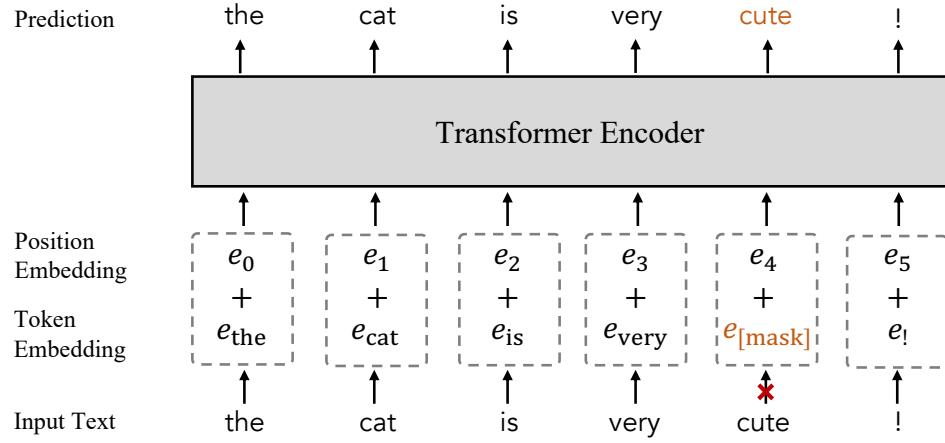


Figure 8: 掩码语言建模的示意图。图引自 [24]。

掩盖约 40% 的图像块（参照 [2]）和 15% 的文本 tokens（参照 [24]），然后利用 MLP 对掩盖的块和 tokens 进行预测，从而捕获丰富的视觉-语言对应信息。

5.2.4 图像到文本生成

图像到文本生成的目标是为给定的图像生成描述性文本，以通过预测 tokens 化的文本来捕获细粒度的视觉-语言相关性。它首先将输入图像编码为中间 embedding 表示，然后解码为描述性文本，如公式 9 所示。例如，COCA[108]、NLIP[44] 和 PaLI[13] 通过标准的编码器-解码器架构和图像描述任务进行训练，如图 9 所示。

5.2.5 讨论

生成目标通过跨模态生成或掩码图像/语言/跨模态建模，促使 VLMs 学习丰富的视觉、语言以及视觉-语言上下文信息，以提高零样本预测能力。因此，生成目标通常作为其他 VLMs 预训练目标之上的附加目标，用于学习丰富的上下文信息 [108, 89, 59]。

5.3 基于对齐目标的 VLMs 预训练

对齐目标通过学习预测给定文本是否正确描述给定图像，促使 VLMs 对齐配对的图像和文本。这一目标可大致分为全局图像-文本匹配和局部区域-词匹配两类。

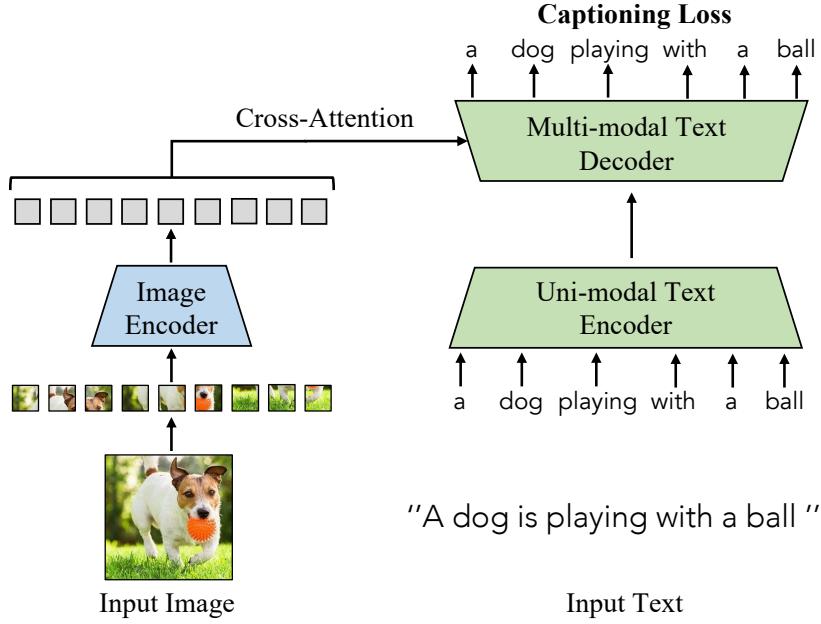


Figure 9: COCA 中图像到文本生成的简化示意图。图引自 [108]。

5.3.1 图像-文本匹配

图像-文本匹配通过直接对齐配对的图像和文本来建模全局的图像-文本相关性，其定义如公式 10 所示。例如，给定一批图像-文本对，FLAVA[89] 使用分类器和二分类损失，将给定图像与其配对的文本进行匹配。FIBER[27] 参考 [3] 的方法，通过成对相似性挖掘难负样本，从而更好地对齐图像和文本。

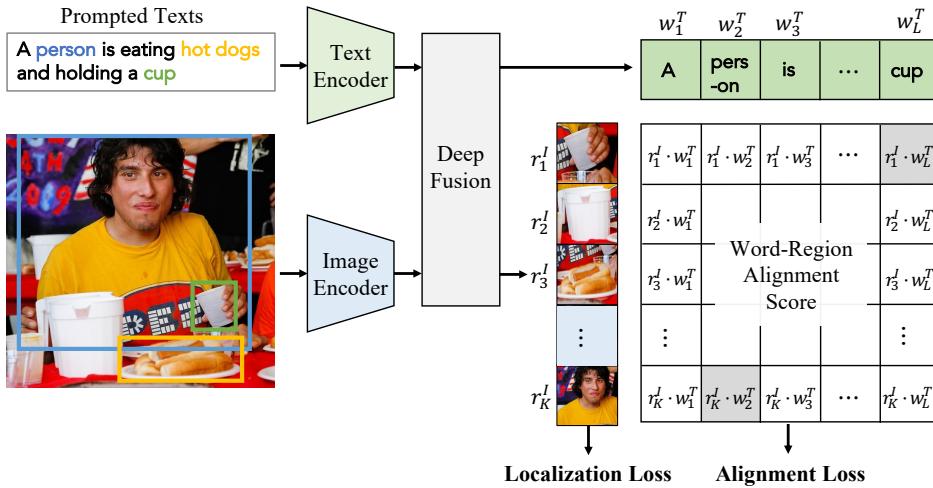


Figure 10: GLIP 使用词-区域对齐进行检测的示意图。图引自 [58]。

5.3.2 区域-词匹配

区域-词匹配目标通过对齐配对的图像区域和词 tokens 来建模局部的细粒度视觉-语言相关性，极大地提升了目标检测和语义分割中的零样本密集预测能力。例如，GLIP[58]、FIBER[27] 和 DetCLIP[105] 用区域-词对齐得分（即区域视觉特征和 tokens 特征之间的点积相似度）替代了对象分类 logits，如图 10 所示。

5.3.3 讨论

对齐目标通过预测图像和文本数据是否匹配来学习跨模态相关性，既简单又易于优化，同时可通过局部匹配图像和文本数据扩展到建模细粒度视觉-语言相关性。另一方面，这些目标通常在单一视觉或语言模态内学习的相关性信息较少。因此，对齐目标通常作为其他 VLMs 预训练目标的辅助损失，用于增强跨视觉和语言模态的相关性建模能力 [89, 116]。

5.4 总结与讨论

综上所述，VLMs 预训练通过不同的跨模态目标（如图像-文本对比学习、掩码跨模态建模、图像到文本生成以及图像-文本/区域-词匹配）来建模视觉-语言相关性。同时，还探索了各种单模态目标以充分挖掘各自模态的数据潜力，例如针对图像模态的掩码图像建模和针对文本模态的掩码语言建模。另一方面，近期 VLMs 预训练专注于学习全局视觉-语言相关性，这对图像分类等图像级别的识别任务有利。与此同时，一些研究 [101, 58, 63, 114, 105, 79, 27] 通过区域-词匹配建模局部细粒度的视觉-语言相关性，以期在目标检测和语义分割中实现更好的密集预测能力。

6 VLMs 预训练的性能比较

如3.4节所述，零样本预测是一种广泛采用的评估设置，用于评估 VLMs 在未见任务上的泛化能力，而无需进行任务特定的微调。本节展示了零样本预测在不同视觉识别任务上的性能，包括图像分类、目标检测和语义分割。

Table 4: 零样本预测设置下 VLMs 预训练方法在图像分类任务上的性能表现。

方法	图像编码器	文本编码器	训练数据量	ImageNet-1k[2]	CIFAR-10[54]	CIFAR-100[54]	Food101[4]	sun397[98]	Cars[52]	Aircraft[64]	DTD[17]	Pets[73]	caltech101[32]	flowers102[70]
CLIP[76]	ViT-L/14	Transformer	400M	76.2	95.7	77.5	93.8	68.4	78.8	37.2	55.7	93.5	92.8	78.3
ALIGN[46]	EfficientNet	BERT	1.8B	76.4	-	-	-	-	-	-	-	-	-	-
OTTER[97]	FBNetV3-C	DeCLUTR-Sci	3M	-	-	-	-	-	-	-	-	-	-	-
DeCLIP[59]	REGNET-Y	BERT	88M	73.7	-	-	-	-	-	-	-	-	-	-
ZeroVL[20]	ViT-B/16	BERT	100M	-	-	-	-	-	-	-	-	-	-	-
FILIP[106]	ViT-L/14	Transformer	340M	77.1	95.7	75.3	92.2	73.1	70.8	60.2	60.7	92.0	93.0	90.1
UniCL[103]	Swin-tiny	Transformer	16.3M	71.3	-	-	-	-	-	-	-	-	-	-
Florence[109]	CoSwin	RoBERT	900M	83.7	94.6	77.6	95.1	77.0	93.2	55.5	66.4	95.9	94.7	86.2
SLIP[69]	ViT-L	Transformer	15M	47.9	87.5	54.2	69.2	56.0	9.0	9.5	29.9	41.6	80.9	60.2
PyramidCLIP[34]	ResNet50	T5	143M	47.8	81.5	53.7	67.8	65.8	65.0	12.6	47.2	83.7	81.7	65.8
Chinese CLIP[102]	ViT-L/14	CNRoberta	200M	-	96.0	79.7	-	-	-	26.2	51.2	-	-	-
LIT[110]	ViT-g/14	-	4B	85.2	-	-	-	-	-	-	-	-	-	-
AltCLIP[15]	ViT-L/14	Transformer	2M	74.5	-	-	-	-	-	-	-	-	-	-
FLAVA[89]	ViT-B/16	ViT-B/16	70M	-	-	-	-	-	-	-	-	-	-	-
KELIP[51]	ViT-B/32	Transformer	1.1B	62.6	91.5	68.6	79.5	-	75.4	-	51.2	-	-	-
COCA[108]	ViT-G/14	-	4.8B	86.3	-	-	-	-	-	-	-	-	-	-
nCLIP[116]	ViT-B/16	Transformer	35M	48.8	83.4	54.5	65.8	59.9	18.0	5.8	57.1	33.2	73.9	50.0
K-lite[86]	CoSwin	RoBERT5	813M	85.8	-	-	-	-	-	-	-	-	-	-
NLIP[44]	ViT-B/16	BART	26M	47.4	81.9	47.5	59.2	58.7	7.8	7.5	32.9	39.2	79.5	54.0
UniCLIP[56]	ViT-B/32	Transformer	30M	54.2	87.8	56.5	64.6	61.1	19.5	4.7	36.6	69.2	84.0	8.0
PaLI[13]	ViT-e	mT5	12B	85.4	-	-	-	-	-	-	-	-	-	-
CLIPPO[95]	ViT-L/16	ViT-L/16	12B	70.5	-	-	-	-	-	-	-	-	-	-
OneR[45]	ViT-L/16	ViT-L/16	4M	27.3	-	31.4	-	-	-	-	-	-	-	-
RA-CLIP[99]	ViT-B/32	BERT	15M	53.5	89.4	62.3	43.8	46.5	-	-	25.6	-	76.9	-
LA-CLIP[30]	ViT-B/32	Transformer	400M	64.4	92.4	73.0	79.7	64.9	81.9	20.8	55.4	87.2	91.8	70.3
ALIP[104]	ViT-B/32	Transformer	15M	40.3	83.8	51.9	45.4	47.8	3.4	2.7	23.2	30.7	74.1	54.8
GrowCLIP[22]	ViT-B/16	Transformer	12M	36.1	60.7	28.3	42.5	45.5	-	17.3	-	71.9	-	23.3

Table 5: 零样本预测设置下 VLMs 预训练方法在图像分割任务上的性能表现。

方法	图像 编码器	文本 编码器	训练 数据量	VOC	PASCAL C.	COCO
				[29]	[68]	[60]
GroupVit[101]	ViT	Transformer	26M	52.3	22.4	-
SegClip[63]	ViT	Transformer	3.4M	52.6	24.7	26.5

表 4展示了 11 个广泛采用的图像分类任务的评估结果。需要注意的是，表中列出了 VLMs 预训练的最佳性能，因为 VLMs 预训练通常具有不同的实现方式。从表 4及图 11可以得出以下三点结论：

1. VLMs 性能通常与训练数据规模密切相关。如图 11左图所示，扩展预训练数据能够带来持续的性能提升；

Table 6: 零样本预测设置下 VLMs 预训练方法在图像检测任务上的性能表现。

方法	图像 编码器	文本 编码器	训练 数据量	COCO [60]	LVIS [38]	LVIS Mini. [38]
RegionClip[114]	ResNet50x4	Transformer	118k	29.6	11.3	-
GLIP[58]	Swin-L	BERT	27.43M	49.8	26.9	34.3
FIBER[27]	Swin-B	RoBERTa	4M	49.3	-	32.2
DetCLIP[105]	Swin-L	BERT	2.43M	-	35.9	-

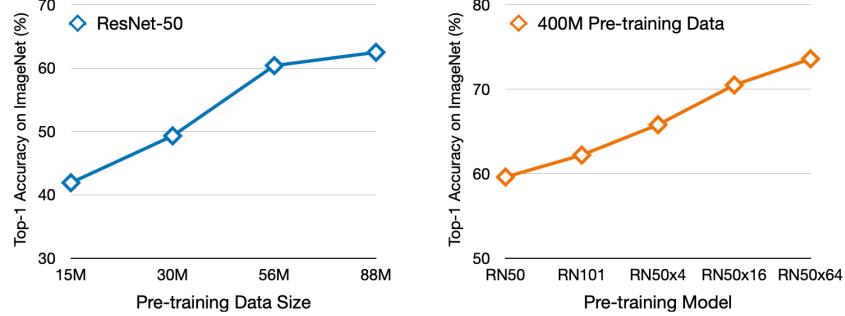


Figure 11: 性能与数据规模及模型规模的关系。结果表明，无论是扩展预训练数据集 [59] 还是扩展预训练模型 [76]，都能显著提升 VLMs 性能。

2. VLMs 性能通常与模型规模密切相关。如图 11 右图所示，在相同的预训练数据条件下，扩展模型规模能够显著提升 VLMs 性能；
3. 通过大规模图文训练数据，VLMs 在多个下游任务中实现了卓越的零样本预测性能。例如，表 4 显示，COCA[108] 在 ImageNet 上取得了最先进的性能，而 FILIP[106] 在 11 个任务中表现稳定。

而 VLMs 卓越的泛化能力主要归因于以下三个因素：

1. 大数据：由于互联网中图文配对数据几乎无限，VLMs 通常使用数百万甚至数十亿的图像和文本样本进行训练，这些数据覆盖了非常广泛的视觉和语言概念，从而具备了强大的泛化能力；
2. 大模型：与传统视觉识别模型相比，VLMs 通常采用更大的模型（例如，COCA 中的 ViT-G 模型 [108] 包含 20 亿参数），这些大模型能够从大数据中高效学习；
3. 任务无关学习：VLMs 预训练中的监督通常是通用的、任务无关的。相比于传统视觉识别中的任务特定标签，图文配对中的文本提供了任务无关、多样且信息丰富的语言监督，帮助训练出适用于多种下游任务的通用模型。

需要指出的是，一些研究 [101, 63, 114, 58, 27, 105] 探索了使用局部 VLMs 预训练目标（如区域-词语匹配 [58]）进行目标检测和语义分割的 VLMs 预训练。表 5 和表 6 总结了这些任务的零样本预测性能。可以观察到，VLMs 在两个密集预测任务中都实现了有效的零样本预测。然而，由于该研究领域仍处于早期阶段且针对密集视觉任务的 VLMs 数量有限，表 5 和表 6 中的结果可能与上述结论不完全一致。

6.1 VLMs 的局限性

尽管前文表明 VLMs 通过扩展数据规模或模型规模能够明显获益，但它们仍然存在以下局限性：

- 随着数据或模型规模的不断增加，性能会出现饱和，进一步扩展不再带来显著提升 [16, 59]；
- 在 VLMs 预训练中使用大规模数据需要巨大的计算资源，例如，在 CLIP 的 ViT-L 模型 [76] 中，需要 256 块 V100 GPU 和 288 小时的训练时间（73728 GPU 时，这是大部分机构无法承担的成本）；
- 使用大规模模型会在训练和推理中引入过高的计算和内存开销。

7 基于预训练模型 VLMs 的技术发展趋势

VLMs 使得能够有效利用网络数据、无需特定任务微调的零样本预测以及对任意类别图像的开放词汇视觉识别。在取得令人瞩目视觉识别性能的同时，VLMs 展现了巨大的潜力。这里，我们列出了七个主要挑战及潜在研究方向

1. 精细化视觉语言相关性建模：通过本地视觉语言对应知识 [58, 105]，VLMs 可以在图像之外更好地识别区域和像素，从而显著提升密集预测任务（如目标检测和语义分割）的效果，这些任务在视觉识别中扮演重要角色。这一方向相关的研究非常有限 [101, 63, 114, 58, 27, 105]。
2. 视觉和语言学习的统一：Transformer 的出现 [96, 26] 使得通过相同方式对图像和文本进行 token 化，可以在单个 Transformer 内统一图像和文本的学习。不同于现有 VLMs[76, 46] 中采用的两个独立网络，统一视觉和语言学习可实现跨模态数据的高效交流，从而提升训练的效果和效率。这一问题已引起一些关注 [95, 45]，但仍需进一步努力以发展更可持续的 VLMs。
3. 多语言 VLMs 预训练：现有 VLMs 多为单语言预训练（如，英语）[76, 46]，这可能导致文化与地域上的偏差 [85, 10]，并限制其在其他语言领域的应用。通过多语言文本进行 VLMs 预训练 [51, 15]，可以学习到不同语言对同一词义的文化视觉特征，从而支持 VLMs 在不同语言环境下的高效应用。
4. 数据高效的 VLMs：大多数现有工作通过大规模训练数据和高强度计算来训练 VLMs，造成可持续性问题。在有限的图文数据下训练高效的 VLMs 可以显著缓解这一问题。例如，与仅从每对图文数据学习相比，可以通过图文数据对之间的监督 [97, 59] 获取更多有用信息。
5. 结合 LLM 的 VLMs 预训练：最近的研究 [30, 104] 利用 LLM 检索丰富的语言知识来增强 VLMs 的预训练。具体而言，它们通过 LLM 扩充原始图文对中的文本，从而提供更丰富的语言知识，帮助更好地学习视觉语言相关性。
6. 结合其他模态的 VLMs：VLMs 在进行预训练时，也存在长尾效应。对于常见物体和生活场景数据几乎是取之不尽；然而，对于工业场景、遥感、医学图像等等专业场景的数据集很有限，并且有个很严重的问题即对于工专业场景数据使用自然语言的描述往往是无力的。此时一些其他模态的数据大概率能够在训练和推理阶段进行很好的辅助。例如图12所示，使用视觉-视觉-语言模型，使用视觉数据进行 Prompt，能够对上述场景问题有一定程度的改善 [47]。
7. 结合图像编辑的 VLMs：基于 NLP 的图像编辑任务（如图13所示）实际上是 VLMs 在视觉任务上的逆问题，这一对任务在训练的过程中应该是相互作用共同进步的关系。更强的图像生成模型例如，Flux[9]，xDiT[31] 等结构有概率能够反作用于 VLMs 对于图像的理解。



Figure 12: 视觉提示 (Visual Prompt) 能够在一定程度上解决 VLMs 在训练时的数据长尾效应和推理时自然语言处理无力描述的问题。图引自 [47]。

References

- [1] Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective. *ArXiv*, abs/2405.16640, 2024.
- [2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014.
- [5] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [6] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.
- [8] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

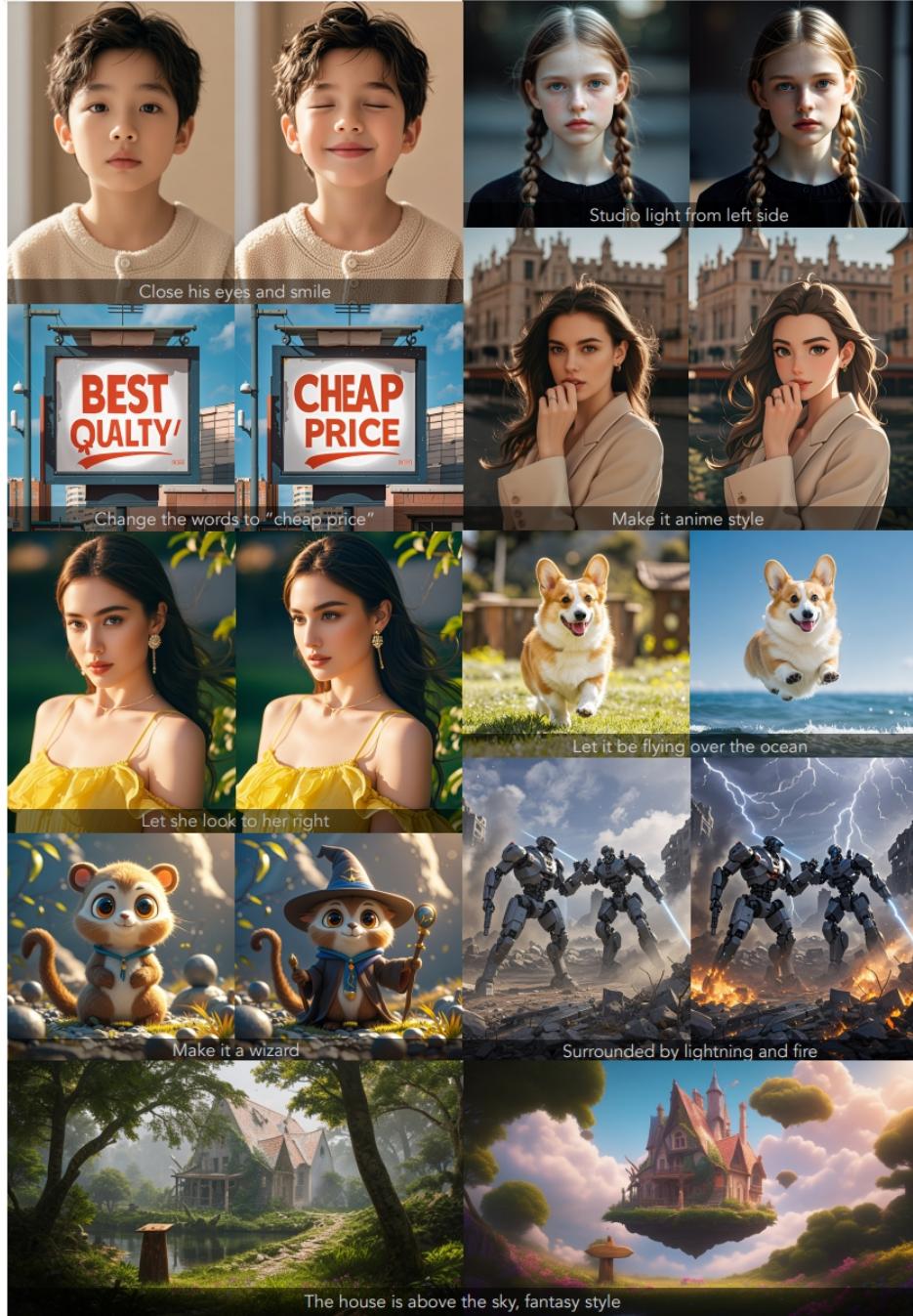


Figure 13: 基于 NLP 的图像编辑任务。图引自 [87]。

- [9] Li-Wen Chang, Wenlei Bao, Qi Hou, Chengquan Jiang, Ningxin Zheng, Yinmin Zhong, Xu-anrun Zhang, Zuquan Song, Chengji Yao, Ziheng Jiang, et al. Flux: Fast software-based communication overlap on gpus through kernel fusion. *arXiv preprint arXiv:2406.06858*, 2024.
- [10] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021.

- [11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.
- [13] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [15] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.
- [16] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023.
- [17] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.
- [18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.
- [19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.
- [20] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *ECCV*, pages 236–253. Springer, 2022.
- [21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.
- [22] Xinch Deng, Han Shi, Runhui Huang, Changlin Li, Hang Xu, Jianhua Han, James Kwok, Shen Zhao, Wei Zhang, and Xiaodan Liang. Growclip: Data-aware automatic model growing for large-scale contrastive language-image pre-training. In *ICCV*, pages 22178–22189, 2023.
- [23] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.
- [24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [25] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022.
- [26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [27] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*.
- [28] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022.
- [29] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

- [30] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023.
- [31] Jiarui Fang, Jinzhe Pan, Xibo Sun, Aoyu Li, and Jiannan Wang. xdit: an inference engine for diffusion transformers (dits) with massive parallelism. *arXiv preprint arXiv:2411.01738*, 2024.
- [32] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004.
- [33] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.
- [34] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramid-clip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.
- [35] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.
- [36] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*, 2022.
- [37] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language distillation. In *ICLR*, 2021.
- [38] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019.
- [39] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.
- [40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.
- [41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [43] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, pages 558–567, 2019.
- [44] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. *arXiv preprint arXiv:2212.07086*, 2022.
- [45] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *AAAI*, volume 37, pages 980–988, 2023.
- [46] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.
- [47] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2025.
- [48] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021.
- [49] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

- [50] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624, 2020.
- [51] Byungsoo Ko and Geonmo Gu. Large-scale bilingual language-image contrastive learning. *arXiv preprint arXiv:2203.14463*, 2022.
- [52] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.
- [53] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.
- [54] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [55] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [56] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *NeurIPS*.
- [57] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022.
- [58] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022.
- [59] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2021.
- [60] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
- [61] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [62] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.
- [63] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2211.14813*, 2022.
- [64] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [65] Ajay Mathur and Giles M Foody. Multiclass and binary svm classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2):241–245, 2008.
- [66] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, 2020.
- [67] Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. *arXiv preprint arXiv:2403.16528*, 2024.
- [68] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014.

- [69] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544. Springer, 2022.
- [70] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008.
- [71] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [72] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011.
- [73] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012.
- [74] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.
- [75] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, pages 647–664. Springer, 2020.
- [76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.
- [77] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [78] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [79] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in vision-language models. *arXiv preprint arXiv:2210.09996*, 2022.
- [80] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.
- [81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [82] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [83] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [84] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019.
- [85] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.
- [86] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022.
- [87] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024.
- [88] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [89] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022.
- [90] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [91] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *ACM SIGIR*, pages 2443–2449, 2021.
- [92] Binyi Su, Hua Zhang, Jingzhi Li, and Zhong Zhou. Toward generalized few-shot open-set object detection. *IEEE Transactions on Image Processing*, 33:1389–1402, 2024.
- [93] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.
- [94] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [95] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Image-and-language understanding from pixels only. *arXiv preprint arXiv:2212.08045*, 2022.
- [96] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.
- [97] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E Gonzalez, and Peter Vajda. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *ICLR*, 2021.
- [98] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.
- [99] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*, pages 19265–19274, 2023.
- [100] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021.
- [101] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022.
- [102] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.
- [103] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, pages 19163–19173, 2022.
- [104] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, pages 2922–2931, 2023.
- [105] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*.
- [106] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021.
- [107] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

- [108] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [109] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- [110] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022.
- [111] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, August 2024.
- [112] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.
- [113] Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, pages 7324–7334. PMLR, 2019.
- [114] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022.
- [115] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.
- [116] Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. *arXiv preprint arXiv:2210.09304*, 2022.
- [117] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.
- [118] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.