自动化学报
ACTA AUTOMATICA SINICA

提示学习在计算机视觉中的分类、应用及展望

刘袁缘 刘树阳 刘云娇 袁雨晨 唐厂 罗威

The Classification, Applications, and Prospects of Prompt Learning in Computer Vision

LIU Yuan-Yuan, LIU Shu-Yang, LIU Yun-Jiao, YUAN Yu-Chen, TANG Chang, LUO Wei

在线阅读 View online: https://doi.org/10.16383/j.aas.c240177

提示学习在计算机视觉中的分类、应用及展望

刘袁缘 [1] 刘树阳 [1] 刘云娇 [1] 袁雨晨 [1] 唐 厂 [1] 罗 威 [2]

摘 要 随着计算机视觉 (CV) 的快速发展, 人们对于提高视觉任务的性能和泛化能力的需求不断增长, 导致模型的复杂度与对各种资源的需求进一步提高. 提示学习 (PL) 作为一种能有效地提升模型性能和泛化能力、重用预训练模型和降低计算量的方法, 在一系列下游视觉任务中受到广泛的关注与研究. 然而, 现有的 PL 综述缺乏对 PL 方法全面的分类和讨论, 也缺乏对现有实验结果进行深入的研究以评估现有方法的优缺点. 因此, 本文对 PL 在 CV 领域的分类、应用和性能进行全面的概述. 首先, 介绍 PL 的研究背景和定义, 并简要回顾 CV 领域中 PL 研究的最新进展. 其次, 对目前 CV 领域中的 PL 方法进行分类, 包括文本提示、视觉提示和视觉−语言联合提示, 对每类 PL 方法进行详细阐述并探讨其优缺点. 接着, 综述 PL 在十个常见下游视觉任务中的最新进展. 此外, 提供三个 CV 应用的实验结果并进行总结和分析, 全面讨论不同 PL 方法在 CV 领域的表现. 最后, 基于上述讨论对 PL 在 CV 领域面临的挑战和机遇进行分析, 为进一步推动 PL 在 CV 领域的发展提供前瞻性的思考.

关键词 计算机视觉, 提示学习, 视觉−语言大模型, 预训练模型

引用格式 刘袁缘, 刘树阳, 刘云娇, 袁雨晨, 唐厂, 罗威. 提示学习在计算机视觉中的分类、应用及展望. 自动化学报, 2025, **51**(5): 1021−1040

DOI 10.16383/j.aas.c240177 **CSTR** 32138.14.j.aas.c240177

The Classification, Applications, and Prospects of Prompt Learning in Computer Vision

LIU Yuan-Yuan[1] LIU Shu-Yang[1] LIU Yun-Jiao[1] YUAN Yu-Chen[1] TANG Chang[1] LUO Wei[2]

Abstract With the rapid development of computer vision (CV), the growing demand for improving the performance and generalization of visual tasks has led to a further increase in model complexity and the need for various resources. Prompt learning (PL), as a method to effectively enhance model performance and generalization, reuse pre-trained models, and reduce computational costs, has gained extensive attention and research in a series of downstream visual tasks. However, existing PL surveys lack comprehensive classification and discussion of PL methods, as well as in-depth analysis of existing experimental results to evaluate the strengths and weaknesses of current methods. Therefore, this paper provides a comprehensive overview of the classification, application, and performance of PL in the field of CV. Firstly, the research background and definition of PL are introduced, followed by a brief review of recent PL progress in CV. Secondly, PL methods in CV are categorized into text prompt, visual prompt, and vision-language joint prompt, with each category elaborated in detail and its strengths and weaknesses discussed. Next, recent advances of PL in ten common downstream visual tasks are reviewed. Additionally, experimental results from three CV applications are provided, summarized, and analyzed to comprehensively discuss the performance of different PL methods in CV. Finally, based on the above discussions, the challenges and opportunities faced by PL in CV are analyzed, offering forward-looking insights to further advance the development of PL in the CV domain.

Key words Computer vision, prompt learning, vision-language large model, pre-trained model

Citation Liu Yuan-Yuan, Liu Shu-Yang, Liu Yun-Jiao, Yuan Yu-Chen, Tang Chang, Luo Wei. The classification, applications, and prospects of prompt learning in computer vision. *Acta Automatica Sinica*, 2025, **51**(5): 1021−1040

近年来, 自然语言处理 (Natural language processing, NLP) 领域普遍采用“预训练+微调”的学习范式, 在多个 NLP 应用任务上取得不错的效果. 该范式首先在大规模数据上进行大语言模型 (Large language model, LLM) 的预训练, 然后在基于特定任务的数据集上进行微调. 随着 LLM 不断

收稿日期 2024-04-04 录用日期 2024-08-27
Manuscript received April 4, 2024; accepted August 27, 2024
国家自然科学基金 (62076227, U2341228), 湖北省自然科学基金 (2023AFB572), 湖北省智能地理信息处理重点实验室 (KLIGIP-2022-B10) 资助
Supported by National Natural Science Foundation of China (62076227, U2341228), Natural Science Foundation of Hubei Province (2023AFB572), and Hubei Key Laboratory of Intelligent Geo-information Processing (KLIGIP-2022-B10)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 中国地质大学 (武汉) 计算机学院 武汉 430074 2. 中国舰船研究设计中心 武汉 430064
1. School of Computer Science, China University of Geosciences, Wuhan 430074 2. China Ship Development and Design Center, Wuhan 430064

发展, 其参数量和对计算资源的需求也呈现出急剧上升的趋势. 以 GPT 系列模型为例, 从 GPT-1 到 GPT-3 的演进中, 其参数量从 1.17 亿激增至 1 750 亿, 同时也带来性能的显著飞跃[1]. 此外, 自 2023 年以来, 一系列 LLMs 被推出, 模型的参数规模也越来越大, 如谷歌的 PaLM 2 (参数量 3 400 亿)、英伟达的 Nemotron-4 (参数量 3 400 亿)、X.AI 的 Grok-1 (参数量 3140 亿) 及华为的 PanGu-Σ (参数量 10 850 亿)[2]. 然而, 这种趋势给“预训练+微调”的学习范式带来巨大的挑战. 首先, 巨大的参数规模使得微调更新预训练模型的所有参数以适应特定的下游任务变得越来越难; 其次, 预训练与下游任务之间存在的域差异性导致大模型难以很好地迁移到其他任务[3].

为解决这些挑战, 一种新的学习范式——提示学习 (Prompt learning, PL) 被提出, 即“预训练+提示+预测”[4]. 在 NLP 领域中, PL 通过在训练样本中添加若干辅助信息, 如精心设计的自然语言, 来帮助预训练大模型适应于特定的下游目标任务. 此过程中, 预训练模型的参数保持冻结, 特定任务所需的标记数据和计算资源需求大大降低. 例如, OpenAI 发布的 ChatGPT, 在无需额外微调的情况下, 只需学习简单的文本提示就能在各种下游任务中表现出卓越的语言理解、处理和生成能力. PL 在 NLP 领域的巨大成功, 吸引了越来越多计算机视觉 (Computer vision, CV) 研究人员的关注和研究.

与 NLP 领域类似, CV 领域的迅猛发展也在很大程度上得益于视觉和多模态基础大模型的发展, 而这些大模型所面临的微调挑战同样显著. 例如, DINOv2[5] 是一个在 1.42 亿张图片上自监督训练的视觉基础模型, 具有超过 10 亿个参数; LVM[1] 是一个具有最高达 30 亿参数的视觉基础模型. 多模态大模型 Yi-VL 使用约一百万个图像−文本对来训练最多 340 亿参数, 在多模态理解与生成任务中表现卓越. 表 1 展示了近年来视觉和多模态大模型的发展及其参数量的变化趋势. 可见, 在 CV 领域, 预训练基础模型也朝着越来越大的方向发展, 而高维的图像和视频数据使得微调这些模型的难度也越来越大. 因此, 如何利用 PL 以提升这些基础大模型在不同 CV 任务上的性能和泛化能力, 降低对标记数据和计算资源的需求成为当前 CV 领域研究的重点. 最近, 已有工作将 PL 成功应用在 CV 领域, 并验证了其可行性. 如 CLIP[6] (Contrastive language-image pretraining) 通过精心设计的文本提示, 不仅在图像分类上表现出色, 还成功应用于多种下游任务中. SAM[7] (Segment anything model) 通过文本或视觉提示能够在无需额外训练的情况下, 灵活执行各种图像分割任务. 随着 PL 在 CV 领域研究的继续深入, 更多创新性的 PL 方法将会涌现, 并不断提升不同下游任务的性能, 推动 CV 领域的发展.

为深入探索 PL 在不同 CV 任务中的工作原理, 廖宁等[8] 对近年的视觉 PL 进行了综述, 讨论在单模态视觉模型和多模态视觉−语言模型中的提示方法及应用. 不同于他们将 PL 方法按照应用模型进行分类, 本文进一步分析不同的 PL 方法本身的模态性质, 将它们分为文本提示、视觉提示和视觉−语言联合提示三个类别. 此外, 本文还在 CV 的多个应用领域讨论不同的 PL 方法, 并在三个广泛应用的 CV 任务上进行结果分析, 充分论证不同 PL 方法的优缺点. 本文对 PL 在 CV 领域的研究和应用进行系统全面的综述. 首先对 CV 领域的发展现状及问题进行论述, 而后分析引入 PL 方法的优势并给出其正式定义. 其次, 按照不同的模态对现有 PL 方法进行分类总结, 并对每种提示方法的原理和优缺点进行分析. 然后梳理 CV 常见下游任务中 PL 的典型应用, 如图 1 所示, 对各种下游任务中的 PL 方法进行分类分析. 接着收集多个代表性 PL 方法和非 PL 方法在具体视觉任务中的实验数据, 并进行详细的对比和分析. 最后探讨当前 PL 在 CV 任务中存在的挑战和潜在的研究方向.

与同类文献相比, 本文的主要贡献如下: 1) 提供一个 PL 在 CV 领域应用的全面系统的综述, 对 PL 的发展过程、基本原理、作用机制进行详细探讨. 根据不同的输入模态, 将现有文献中的 PL 方法分类为文本提示、视觉提示和视觉−语言联合提示, 并分别讨论它们的基本原理和优缺点. 此外, 本文系统地讨论 PL 在 CV 领域下游任务中的典型应用, 根据提示类型阐述不同 PL 方法在具体下游任务中的作用机制. 2) 针对多个 CV 下游任务进行实验数据的收集和定量分析, 对比代表性 PL 方法与

表 1 CV 领域视觉与多模态基础大模型及其参数量

Table 1 Vision and multimodal foundational large models in CV with their parameter size

	视觉模型				多模态模型				
	DERT	Vision Transformer	DINOv2	LVM	CLIP	SAM	MiniGPT-4	LLaVA	Yi-VL
年份	2020	2021	2023	2023	2021	2023	2023	2023	2024
参数量	40 M	86 M ~ 632 M	1.1 B	300 M ~ 3 B	400 M ~ 1.6 B	1 B	13 B	7 B ~ 13 B	6 B ~ 34 B

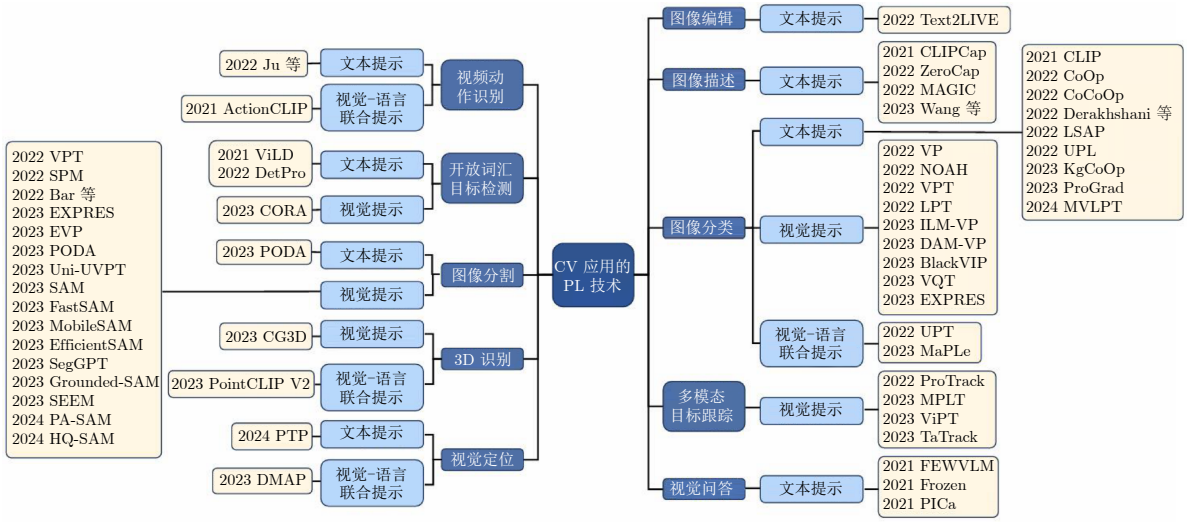

图 1　基于 PL 的 CV 应用概述

Fig. 1　Overview of CV applications based on PL

非 PL 方法之间的性能差异以及需训练的参数量的差异, 从量化的角度探讨 PL 方法的优异性. 3) 基于 CV 领域 PL 方法的基本原理讨论现有 PL 方法和发展现状, 详细讨论现有 PL 方法应用于 CV 任务时面临的挑战和机遇, 对 PL 方法在 CV 领域的未来发展提供一些见解, 为未来探索新颖的 PL 方法以及进一步促进不同类型的 PL 方法的交互与融合提供指引.

1　提示学习

1.1　自然语言处理中的提示学习

为了更好地利用预训练模型蕴含的知识, 在 NLP 领域首先提出了 PL, 与传统监督学习不同, PL 不需要直接对模型的参数进行训练, 而是在文本输入 x 的基础上, 应用提示函数 $f_{prompt}(\cdot)$ 将 x 建模为文本提示, 再使用该提示预测输出 y, 实现了少样本甚至零样本学习. NLP 中的基本提示流程如下:

1) 添加提示. 使用提示函数 $f_{prompt}(\cdot)$ 将文本输入 x 建模为文本提示 x',

$$x' = f_{prompt}(x) \tag{1}$$

具体来说, 提示函数通常使用一个文本字符串模板, 该模板包含两个填充槽, 一个输入槽 $[X]$ 用于填充输入 x, 一个答案槽 $[Z]$ 用于填充潜在答案 z.

2) 答案搜索与映射. 首先, 将 Z 定义为 z 的允许取值集合, 将提示中的位置 $[Z]$ 用潜在答案 z 填充后, 输入预训练语言模型 $P(\cdot;\theta)$ 中, 通过评估潜在答案集合对应的填充提示在模型中的表现来获得答案. 再将对应答案映射到输出值 y.

$$z' = search_{z\in Z} P(f_{fill}(x',\ z);\ \theta) \tag{2}$$

其中, f_{fill} 表示使用不同潜在答案填充的提示, $search_{z\in Z}$ 可以是 argmax 等函数.

例如, 定义 Z 为 {出色的, 良好的, 一般的, 糟糕的, 很差的}, 来表示标签集 Y 中每个情感标签类 {++, +, ∼, −, −−}, 将包含在 Z 中的 z 填充进模板后, 输入预训练模型中, 使用搜索函数得到最匹配的答案 z', 最后再将 z' 映射到对应的输出 y[4]. NLP 中的提示流程如图 2 所示.

1.2　计算机视觉中提示学习的分类

在 CV 领域, 受益于深度学习和大规模数据集的发展, 引入额外的提示信息指导模型学习已成为一种行之有效的方法. 根据 PL 方法本身的数据模态不同, 将现有的提示方法划分为文本提示、视觉提示和视觉–语言联合提示.

1.2.1　文本提示

文本提示已被广泛应用于视觉–语言模型中, 旨在引导模型更好地提取和理解文本信息, 弥合多模态之间的差距, 并指导模型有效地处理和生成与视觉输入相关的响应. 如图 3 所示, 本文将 CV 任务中常见的文本提示分为基于手工设计的文本提示、连续提示、基于梯度引导的文本提示、基于视觉映射到语言空间的提示、基于图像引导的文本提示、基于伪标签的文本提示、基于多任务的文本提示.

基于手工设计的文本提示. 主要通过设计合适的提示模板对输入标签或文本进行处理, 来实现视觉与文本特征更好的对齐, 如图 3(a) 所示. 然而, 设计合适的提示模板是一项关键且具有挑战性的任

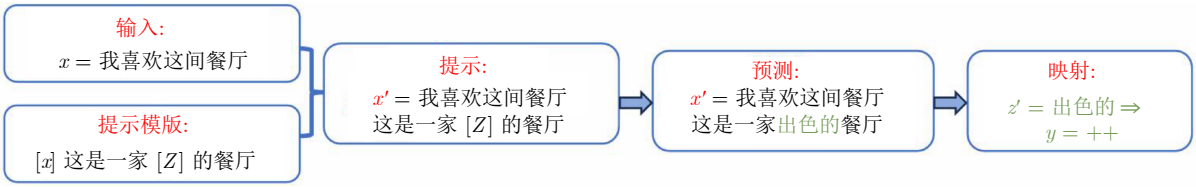

图 2 NLP 中的提示流程

Fig. 2 The prompting process in NLP

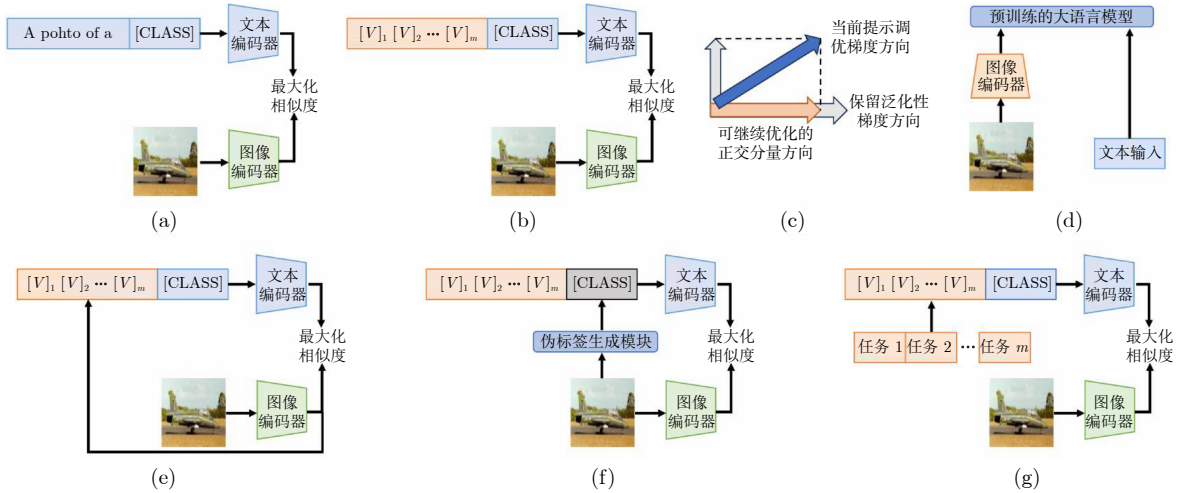

图 3 文本提示 ((a) 基于手工设计的文本提示; (b) 连续提示; (c) 基于梯度引导的文本提示; (d) 基于视觉映射到语言空间的提示; (e) 基于图像引导的文本提示; (f) 基于伪标签的文本提示; (g) 基于多任务的文本提示)

Fig. 3 Text prompts ((a) Text prompt based on hand-crafted; (b) Continuous prompt; (c) Text prompt based on gradient guidance; (d) Prompt based on the mapping from vision to the language space; (e) Text prompt based on image guidance; (f) Text prompt based on pseudo-labels; (g) Text prompt based on multi-task)

务. 针对特定任务场景的固定提示模板往往缺乏通用性, 且细微的变化都可能对最终性能产生不可预料的影响. 因此, 设计提示模板往往需要反复地实验对比, 耗时耗力且难以保证其有效性和通用性.

连续提示. 为了提升提示模板的获取效率和性能, 连续提示被提出以自动在文本嵌入空间中搜索提示模板. 如图 3(b) 所示, 连续提示通常使用一组可学习的提示向量来代替手工设计的提示模板. 通过在特定下游任务的训练数据上进行提示调优, 从而实现了提示向量的自动化生成. 相比于基于手工设计的文本提示, 连续提示牺牲了一定的泛化能力, 但是获得了更好的闭集识别性能.

基于梯度引导的文本提示. 为了解决连续提示在调优中可能出现的泛化能力下降问题, 如图 3(c) 所示, 基于梯度引导的文本提示确保在训练过程中提示调优的方向与预训练模型中固有知识的通用方向不产生冲突, 从而避免在学习过程中对模型泛化能力产生不利影响.

基于视觉映射到语言空间的提示. 为了将单模态 LLMs 应用于视觉任务, 一种基于视觉映射到语言空间的提示方法被提出. 如图 3(d) 所示, 此方法使用视觉编码器将视觉输入映射到语言空间中, 作为 LLMs 能够理解和处理的文本输入. 此方法充分利用了 LLMs 在语言理解和生成方面的强大能力, 使模型能够在下游 CV 任务中实现准确的分类和推理.

基于图像引导的文本提示. 在视觉−语言模型中, 仅针对特定下游任务优化文本端提示参数的做法往往无法充分地捕获和利用视觉输入中的有效信息, 导致模型的泛化能力不足. 为了弥补该缺陷, 如图 3(e) 所示, 基于图像引导的文本提示将处理过的视觉特征整合到文本空间, 并通过引导文本端提示参数的更新来实现两种模态更好的融合与对齐.

基于伪标签的文本提示. 为了解决真实世界中目标数据标签稀缺的问题, 一种基于伪标签的文本提示方法被提出. 如图 3(f) 所示, 此方法通过使用一个伪标签生成模块为目标数据生成伪标签, 而后使用伪标签数据训练一组提示向量. 此方法充分利用预训练的伪标签生成模块中存储的先验知识, 能有效解决标签稀缺问题.

基于多任务的文本提示. 为了充分发掘和利用

不同视觉任务之间的联系, 如图 3(g) 所示, 基于多任务的文本提示提出让多个任务共享一组提示参数来实现多任务学习. 只需要引入少量提示参数, 即可有效地在一个通用模型上完成多任务处理.

1.2.2　视觉提示

如图 4 所示, 视觉提示方法可分为: 基于像素扰动的视觉提示、基于提示 tokens 的视觉提示、基于提示模块的视觉提示、基于上下文样例模板的视觉提示和基于网络结构搜索的视觉提示. 总体上, 各类提示机制通过重整下游任务特征, 使其与预训练模型的匹配程度更高, 达到重用预训练模型的目的.

基于像素扰动的视觉提示. 基于像素扰动的视觉提示在输入图像上附加额外的可学习像素扰动, 可以细分为共享像素扰动提示和特定像素扰动提示, 如图 4(a) 所示. 共享像素扰动提示被整个下游任务样本共享, 特定像素扰动提示 P_i 负责处理特定的样本组. 像素扰动提示中新引入的提示参数基于反向传播优化.

基于提示 tokens 的视觉提示. 提示 tokens 在视觉 Transformer 编码层的输入序列中引入可学习的 tokens, 可以细分为共享提示 tokens 和特定提示 tokens, 如图 4(b) 所示. 共享提示 tokens 被整个下游任务样本共享, 特定提示 tokens P_i 负责捕捉特定的特征信息. 在前向传播过程中, 提示 tokens 与图像 tokens 和类别 tokens 通过注意力机制交互, 转换下游任务的特征, 使其匹配预训练模型的特征表示, 辅助预训练模型在下游任务上产生更准确的预测结果.

基于提示模块的视觉提示. 轻量级神经网络模块作为提示模块, 被额外添加到预训练模型的不同位置, 可以细分为添加在主干网络内部的提示模块和主干网络外的提示模块, 如图 4(c) 所示. 利用预训练模型解决下游任务时, 保持主干网络参数不变, 只优化提示模块的参数.

基于上下文样例模板的视觉提示. 将原始输入图像和对应的预测结果作为样例提示, 进一步与待预测的原始图像和空白的占位图像拼接起来, 形成上下文样例模板提示, 如图 4(d) 所示. 在上下文样例模板提示的引导下, 模型能够输出预测结果, 填充空白占位图像.

基于网络结构搜索的视觉提示. 现有的视觉提示方法如提示 tokens、提示模块等组成网络结构搜索空间, 基于网络结构搜索的视觉提示在该搜索空间中随机选择一种现有的视觉提示方法, 与预训练模型组合并调优, 如图 4(e) 所示. 对不同的提示方法进行性能对比, 从中选取最优的提示方法.

1.2.3　视觉–语言联合提示

视觉–语言联合提示将上述文本提示和视觉提示相结合, 同时引入视觉–语言模型中, 以促进两种模态的特征实现更好的交互和融合. 已有工作, 如

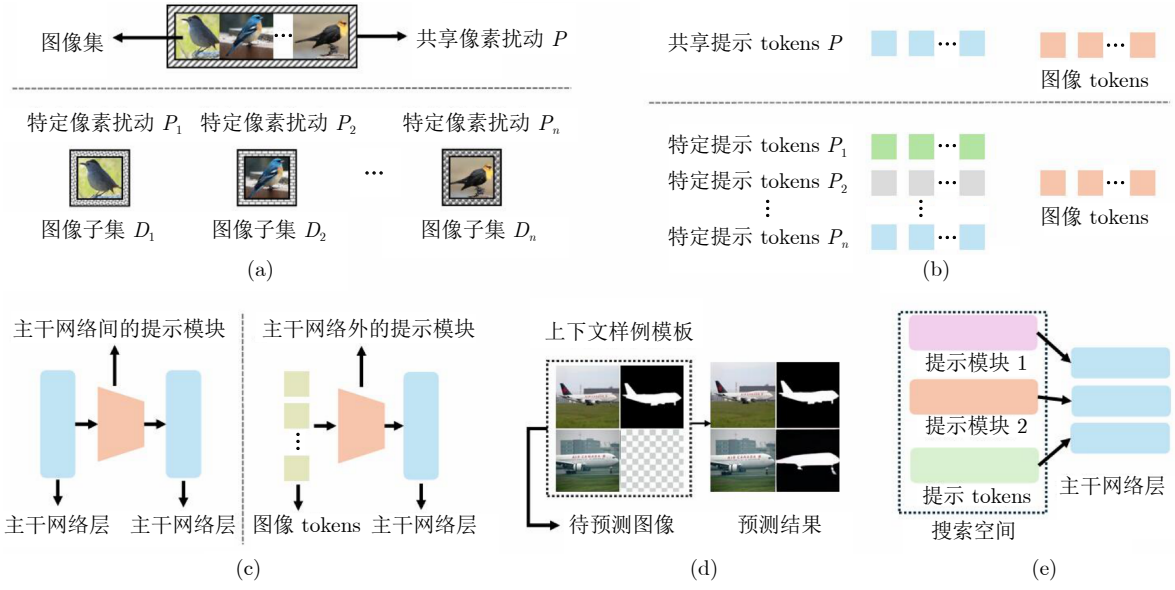

图 4　视觉提示 ((a) 基于像素扰动的视觉提示; (b) 基于提示 tokens 的视觉提示; (c) 基于提示模块的视觉提示; (d) 基于上下文样例模板的视觉提示; (e) 基于网络结构搜索的视觉提示)

Fig. 4　Visual prompts ((a) Pixel perturbation-based visual prompt; (b) Prompt tokens-based visual prompt; (c) Prompt module-based visual prompt; (d) Contextual example template-based visual prompt; (e) Network architecture search-based visual prompt)

UPT[9] 和 MaPLe[10], 证明了在视觉–语言模型中引入多模态的视觉–语言联合提示能够进一步发掘两种提示的潜力, 有效提升模型的性能和泛化能力.

如图 5 所示, 其展示了在视觉–语言模型上四种联合提示交互和融合的方法. 图 5(a) 中两种模态独立学习各自的可学习提示向量, 图 5(b) 中训练两种模态共享的提示 tokens, 图 5(c) 中使用两个全连接层来生成两种模态的提示, 图 5(d) 中使用一个 Transformer 层来生成两种模态的提示. 相关实验结果表明, 图 5(d) 中使用同一个 Transformer 层来生成两种提示, 这种方法能够更好地提升两种模态的交互性, 在 11 个图像分类数据集上获得了优于单模态提示的性能.

2 计算机视觉中提示学习的应用

近年来, 通过 PL 辅助预训练模型能够更有效地解决各种下游视觉任务, 如图像分类、图像分割、开放词汇目标检测、视频动作识别、图像描述、多模态目标跟踪、视觉问答、视觉定位、3D 识别、图像编辑等已经取得了一定的进展. 下面将详细介绍不同视觉任务中的代表性 PL 方法, 并分析其优缺点.

2.1 图像分类任务

图像分类[11] 是 CV 领域的核心任务之一, 旨在训练模型以自动识别和区分不同类别的图像. 面向图像分类任务的 PL 方法主要包括基于文本提示的图像分类方法、基于视觉提示的图像分类方法和基于视觉–语言联合提示的图像分类方法.

2.1.1 基于文本提示的图像分类方法

通过在视觉–语言模型的文本端引入不同的提示方法来辅助模型提取到更加准确的文本特征, 促进文本特征与视觉特征之间实现更好的对齐, 来拉近视觉特征与对应类别的文本特征之间的距离, 实现有效的视觉识别.

不同于以往的模型直接学习输入图像与标签之间的对应关系, CLIP[6] 使用一个手工设计的提示模板将输入标签处理为一个完整的自然语言句子, 如 “A photo of [CLASS]”作为文本提示被提取为文本特征, 随后与图像编码器从输入图像中提取的视觉特征计算相似度. 通过引入这种手工设计的提示, CLIP 模型能够在没有任何额外微调的情况下, 在 ImageNet[12] 分类数据集上展现出出色的零样本/少样本识别能力.

为了进一步提升基于手工设计的文本提示方法的性能, Zhou 等[13] 在 CLIP 的基础上提出基于连续提示的图像分类模型 CoOp (Context optimization). 通过使用一组可学习的提示向量取代固定的提示模板, 并通过反向传播进行调优, CoOp 实现了对提示模板的自动学习而非手动设计. 虽然引入了新的提示参数并损失了一定的泛化能力, 但在 11 个典型的图像分类数据集上, CoOp 的准确率相较于 CLIP 都有了明显的提升.

为了解决连续提示导致泛化能力下降的问题, Zhou 等[14] 使用基于图像引导的文本提示提出 CoCoOp (Conditional context optimization). CoCoOp 引入一个轻量级神经网络 (Meta-net), 为每个图像生成基于视觉输入的 tokens, 并整合到文本端的提示向量中, 从而产生适应每个实例且对类别迁移不敏感的动态提示. 这种提示方法使得 CoCoOp 在基类上的性能达到 CoOp 的水平, 而在泛化能力上接近 CLIP. 类似地, Derakhshani 等[15] 使用图像编码器从输入图像中提取图像特征, 并对这些特征进行高斯分布建模, 而后从该分布中随机采样一个特征向量与文本端的提示向量相结合, 形成

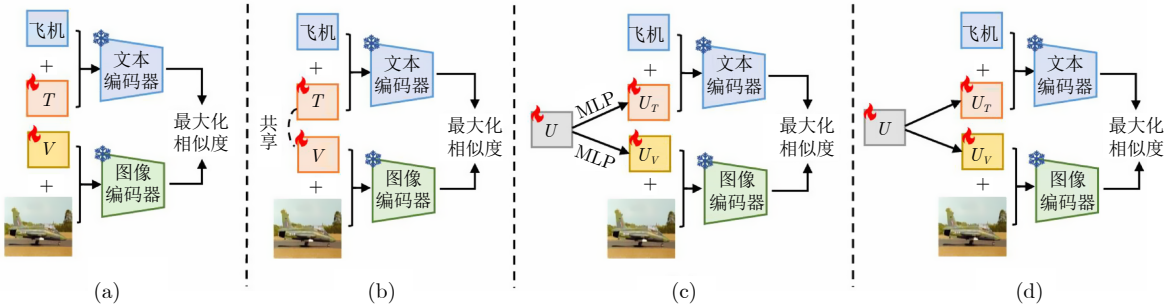

图 5 在视觉–语言模型上引入视觉–语言联合提示的四种方法对比 ((a) 独立训练两种模态的提示; (b) 共享地训练两种模态的提示; (c) 使用两个 MLP 层来生成提示; (d) 使用一个轻量级的自注意力网络来生成提示)

Fig. 5 Comparison of four methods for introducing vision-language joint prompts in vision-language models ((a) Independently train the prompts of the two modalities; (b) Train the prompts of two modalities in a shared manner; (c) Utilizing two MLP layers to generate prompts; (d) Employing a lightweight self-attention network to generate prompts)

基于图像引导的文本提示来引导文本提示更好地学习. 通过这种方式, 模型能够在保持对基类识别能力的同时增强对新类的泛化能力.

为了保留基于手工设计的文本提示方法的泛化能力和连续提示方法的闭集识别精度, KgCoOp[16] 同时引入了这两种提示, 并通过增加损失约束来最大限度地降低由两种提示方法生成的文本嵌入之间的差异. 相较于 CoCoOp, 此模型在基类和新类上的分类精度都有了进一步提升. 类似地, LASP[17] 在此基础上设置多组基于手工设计的文本提示和连续提示, 通过降低同组的两种提示方法生成的文本嵌入之间的差异, 模型能够更细致地捕捉到不同类别之间的细微差异, 在闭集分类精度和泛化能力上都超越了 CLIP 和 CoOp.

Zhu 等[18] 发现在样本稀缺的情况下, 基于连续提示的 CoOp 模型的泛化能力可能会随着训练的进行而下降, 甚至会低于零样本预测. 这表明在对连续提示优化的过程中存在灾难性遗忘问题. 为了在训练过程中保护模型的泛化能力, 他们提出一种基于梯度引导的文本提示的 ProGrad 方法, 专注于优化提示参数的微调方向. ProGrad 通过计算预训练模型的零样本预测与微调模型预测之间的 KL 散度来界定一般知识方向, 并利用特定任务的交叉熵损失梯度来细化这一方向. 只有在更新方向与一般知识相一致或不冲突的情况下, 模型才会沿此方向进行微调. ProGrad 有效防止了在微调过程中对预训练知识的遗忘, 从而增强了模型的泛化能力.

为了解决数据缺乏标签的问题, UPL[19] 使用基于伪标签的文本提示来引入伪标签并优化目标提示. 具体而言, 首先使用 CLIP 模型和一个基于手工设计的文本提示“A photo of a [CLASS]”来为无标签的目标数据生成伪标签, 并为每个类挑选出置信度最高的 Top-K 个样本, 而后利用这些挑选出的高置信度样本来优化连续提示. 通过这一提示策略, UPL 有效利用了未标注数据, 增强了视觉–语言模型在无监督学习环境下的性能和泛化能力.

为了解决现有提示方法在单任务上表现优异, 但无法有效迁移到多任务场景的问题, MVLPT[20] 应用基于多任务的文本提示方法, 首先从多个源任务中学习一组通用的共享提示向量来捕捉不同任务之间的共性, 并以此来初始化不同目标任务的提示. 而后根据不同目标任务之间的相关性进行分组, 在各个组内通过提示调优来适应目标任务. MVLPT 有效地将跨任务知识融入到视觉–语言模型的提示调优中, 显著增强了模型在面对多任务时的泛化能力和适应性.

2.1.2　基于视觉提示的图像分类方法

VP[21] (Visual prompting) 引入基于共享像素扰动的视觉提示, 添加到图像的边框位置, 在下游任务上微调时只优化引入的提示参数. 类似地, ILM-VP[22] (Iterative label mapping-based visual prompting) 也在图像的边框位置添加基于共享像素扰动的视觉提示. BlackVIP[23] 为每一张输入图片生成一个基于特定像素扰动的视觉提示, 添加到整张图像上. DAM-VP[24] (Diversity-aware meta visual prompting) 为数据集中的不同子集分配一个基于特定像素扰动的视觉提示. 如图 6(a) 所示, DAM-VP 基于现有的聚类方法划分子集, 为每个子集单

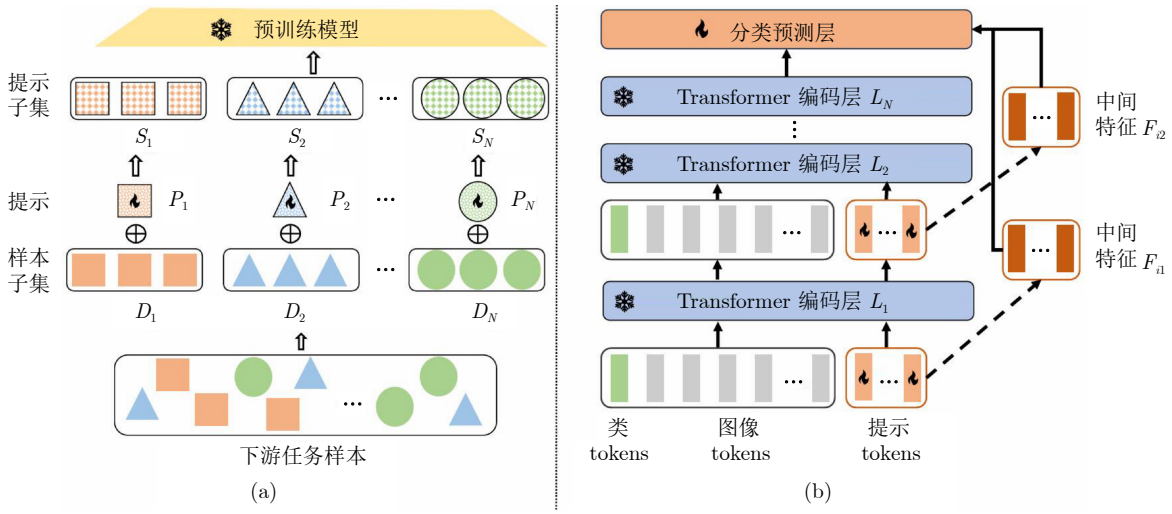

图 6　图像识别中的视觉提示方法 ((a) 基于像素扰动提示的 DAM-VP; (b) 基于提示 tokens 的 VQT)

Fig. 6　Visual prompt methods in image recognition ((a) DAM-VP based on pixel perturbation prompts; (b) VQT based on prompt tokens)

独优化一份提示.

VPT[25] (Visual prompt tuning) 将基于提示 tokens 的视觉提示引入 Transformer 的一个或多个编码层, 分别对应于 VPT-Shallow 和 VPT-Deep. VQT[26] (Visual query tuning) 与 VPT-Deep 类似, 在 Transformer 的每一层都添加提示 tokens, 但这些 tokens 在注意力交互阶段只参与 query 的计算, 不改变在下游任务上提取的中间特征, 只起到特征聚合的作用, 最终的线性分类层利用提示 tokens 逐层产生的聚合中间特征生成分类结果, 如图 6(b) 所示. EXPRES[27] (Expressive prompts with residuals) 引入基于特定提示 tokens 的视觉提示, 包含浅层提示和残差提示. 浅层提示添加到 Transformer 的输入层, 负责与图像 tokens 交互. 残差提示添加到浅层提示的对应输出位置, 逐层增强 tokens 之间的关联. 类似地, LPT[28] (Long-tailed prompt tuning) 引入基于特定提示 tokens 的视觉提示, 包含组共享提示和组特定提示, 分别用于学习所有类别的共享特征和具有类别区分度的特征, 从而将预训练模型迁移到长尾目标分类任务.

NOAH[29] 采用基于网络结构搜索的视觉提示, 将多种 PL 方法如 VPT、LoRA[30]、Adapter[31] 作为网络结构搜索空间. 对于每一个下游任务, NOAH 从搜索空间中选取其中一个子模块添加到预训练模型中, 在下游任务上分别优化, 选取性能最佳的子模块作为最终提示.

2.1.3 基于视觉–语言联合提示的图像分类方法

尽管现有的添加单模态 (文本或者视觉) 提示的方法取得了广泛的成果, 但 Zang 等[9] 研究发现这些方法在不同数据集上无法保证始终获得高性能, 存在一定的局限性. 例如, CoOp 模型 (基于单模态的文本提示) 在 Flowers102[32] 数据集上的表现优于 VPT 模型 (基于单模态的视觉提示), 而 VPT 模型在 EuroSAT[33] 数据集上表现更好. 因此, 本文尝试在视觉–语言模型的两端同时引入提示以更好地对齐两种模态的特征, 并进一步提升模型的分类精度.

Khattak 等[10] 将视觉–语言联合提示方法和深度提示方法相结合, 提出一种名为 MaPLe (Multimodal prompt learning) 的方法. 如图 7 所示, MaPLe 同时在模型的视觉编码器和语言编码器中引入了可学习的提示向量, 以实现两种模态更好的对齐. 此外, MaPLe 还提出一种深度提示策略, 使提示被整合到两个编码器的前若干个 Transformer 块中, 进一步提升了提示方法的灵活性. 此外, UPT[9] 在 CLIP 上对比和分析了四种不同的联合提

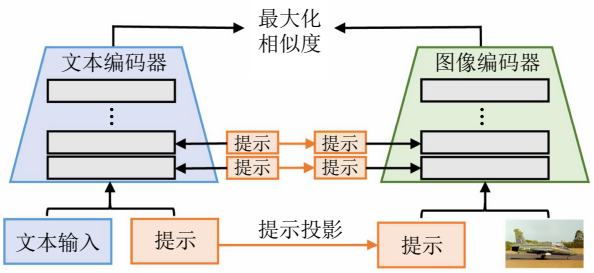

图 7 基于视觉–语言联合提示的 MaPLe 图像分类框架

Fig. 7 Vision-language joint prompts-based MaPLe image classification framework

示方法 (见第 1.2.3 节), 在 11 个分类数据集上的实验结果表明这种联合提示的方法有效地提升了模型的准确性和泛化性能.

2.2 图像分割任务

图像分割[34–46] 是 CV 的基础任务之一, 旨在区分图像中不同区域的像素, 从而实现对图像的细粒度理解和区分. 面向图像分割任务的 PL 方法主要包括基于文本提示的图像分割方法和基于视觉提示的图像分割方法.

2.2.1 基于文本提示的图像分割方法

Fahes 等[34] 提出的 PODA (Prompt-driven zero-shot domain adaptation) 使用基于手工设计的文本提示解决零样本域自适应分割问题, 将对目标域的描述性语句作为文本提示, 生成与目标域风格一致的虚拟特征, 微调模型实现零样本迁移.

2.2.2 基于视觉提示的图像分割方法

VPT[25] 和 EXPRES[27] 在视觉 Transformer 的每一层引入基于提示 tokens 的视觉提示, 同步优化提示参数与分割头部参数. Liu 等[35] 设计基于提示模块的视觉提示 SPM (Semantic-aware prompt matcher), 插入预训练主干网络的相邻阶段之间. SPM 逐阶段生成中间语义图, 处理提取的目标域特征, 使其匹配预训练模型每一阶段的期望输入. 类似地, Liu 等[36] 引入基于提示模块的视觉提示 EVP (Explicit visual prompting), 用于学习输入图像的 patch embeddings 和图像高频部分的融合表示, 利用图像的高频部分为预训练的分割模型提供额外的监督信息. Bar 等[37] 构建基于上下文样例模板的视觉提示, 引导模型根据上下文补全提示中的空白占位图像. 将示例图像、示例分割掩码、待预测目标图像和空白的占位图像拼接形成一张组合图像, 作为上下文样例模板, 引导模型产生预测结果. Uni-UVPT[38] (Universal unsupervised visual prompt tuning) 引入基于提示模块的视觉提示, 插入冻结

的源域预训练模型中, 用于解决域自适应语义分割任务. 提示模块转换目标域特征, 使其逐阶段匹配源域冻结模型的特征.

NLP 领域预训练的基础模型 ChatGPT 能够通过输入提示的引导执行多种不同的下游任务, 为了实现类似的功能, Meta AI 开发了一个图像分割领域的交互式分割基础大模型 SAM[7]. SAM 能够根据输入提示的引导在无需额外微调的情况下执行语义分割、实例分割、全景分割等多种分割任务. SAM 模型基于前景/背景点、粗粒度的标注框或掩码、文本等提示返回图像中相应对象的分割掩码, 如图 8 所示, SAM 的图像编码器和提示编码器分别产生图像 embeddings 和提示 embeddings, 进一步输送到掩码解码器输出对象掩码. 尽管 SAM 分割能力出众, 但是其图像编码器的参数量高达 632 M, 执行分割任务时显存占用多、推理时间长. 为进一步优化, FastSAM[39]、MobileSAM[40]、EfficientSAM[41] 使用不同的方法对 SAM 原有的图像编码器进行轻量化处理, 使用类似 SAM 的点、框、掩码或文本提示产生分割掩码. SAM 对于物体细节部分和特殊形状的物体产生的掩码预测存在边界粗糙问题. 为了提高掩码质量, HQ-SAM[42] 在 SAM 原始的掩码解码器前引入一些提示 tokens, 引导解码器产生精细化掩码; PA-SAM[43] 引入提示模块辅助掩码的产生, 提示模块利用掩码解码器的中间特征产生具有细节信息的 tokens, 以残差的形式融入掩码解码器来优化中间特征表示. SAM 利用输入的点、框等交互式提示在待预测图像上分割出对应物体, 而 SegGPT[44] 使用示例图像及图像中的物体掩码作为提示, 引导模型在待预测图像上分割类似物体.

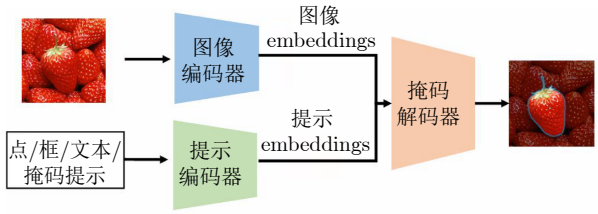

图 8　SAM 方法流程图

Fig. 8　Flowchart of the SAM method

另外, SegGPT 可以优化一个可学习的像素扰动提示作为任务导向, 提高模型对特定任务的分割能力. Grounded-SAM[45] 使用目标检测模型 Grounding DINO 产生的物体定位框作为提示, 实现零样本分割. SEEM[46] 能够依据文本提示、视觉提示、交互式提示或提示组合产生对应的分割掩码, 克服了 SAM 只支持单一类型提示的缺点. 由于交互式分割通常不能一次性完成, SEEM 额外引入提示 tokens 存储历史分割信息, 用于持续细化分割掩码.

2.3　开放词汇目标检测任务

开放词汇目标检测[47−49] (Open-vocabulary object detection, OVD) 是一种特殊的目标检测任务, 与传统目标检测相比, OVD 需要同时关注训练集中的已知类别和现实世界中不断出现的未知类别. 面向开放词汇目标检测任务的 PL 方法主要包括基于文本提示的开放词汇目标检测方法和基于视觉提示的开放词汇目标检测方法.

2.3.1　基于文本提示的开放词汇目标检测方法

ViLD[47] 引入基于手工设计的文本提示模板, 与不同类别组合形成完整的文本提示, 输入 CLIP 的文本编码器产生文本特征, 与检测框对应的区域特征进行对比学习, 产生最终的分类结果. 由于基于手工设计的文本提示存在难以设计和性能次优等问题, Du 等[48] 提出的 DetPro (Detection prompt) 效仿 CoOp, 将提示模板用一些可学习的提示 tokens 代替, 与不同的类别 tokens 组合形成文本提示, 而后这些文本提示被用作区域分类器来指导目标检测器完成训练. 图 9(a) 展示了 ViLD 和 DetPro 的提示机制.

2.3.2　基于视觉提示的开放词汇目标检测方法

CORA[49] 在 CLIP 的图像编码器中引入基于提示 tokens 的视觉提示作为区域提示, 用于增强检测框对应的区域特征, 以减轻整体图像特征与裁剪的区域特征之间的分布差异, 进一步利用 CLIP 模型的对比学习实现区域分类, 如图 9(b) 所示.

2.4　视频动作识别任务

近年来, 视频动作识别[50−51] 取得了显著的进展. 该任务通过分析视频帧的时间信息和视觉内容, 自动识别和分类视频中执行的动作或活动. 面向视频动作识别任务的 PL 方法主要包括基于文本提示的视频动作识别方法和基于视觉–语言联合提示的视频动作识别方法.

2.4.1　基于文本提示的视频动作识别方法

Ju 等[50] 在 CLIP 的基础上使用基于多任务的文本提示将三种视频相关任务转换为与训练目标相同的格式. 具体而言, 模型向文本编码器中添加多个可学习的提示向量, 以生成动作分类器或查询嵌入; 在 CLIP 图像编码器之后应用一个轻量级的 Transformer 进行时序建模. 在训练过程中, 图像和文本编码器都保持冻结状态. 通过优化特定任务的提示向量和时序 Transformer, 有效地将 CLIP 适

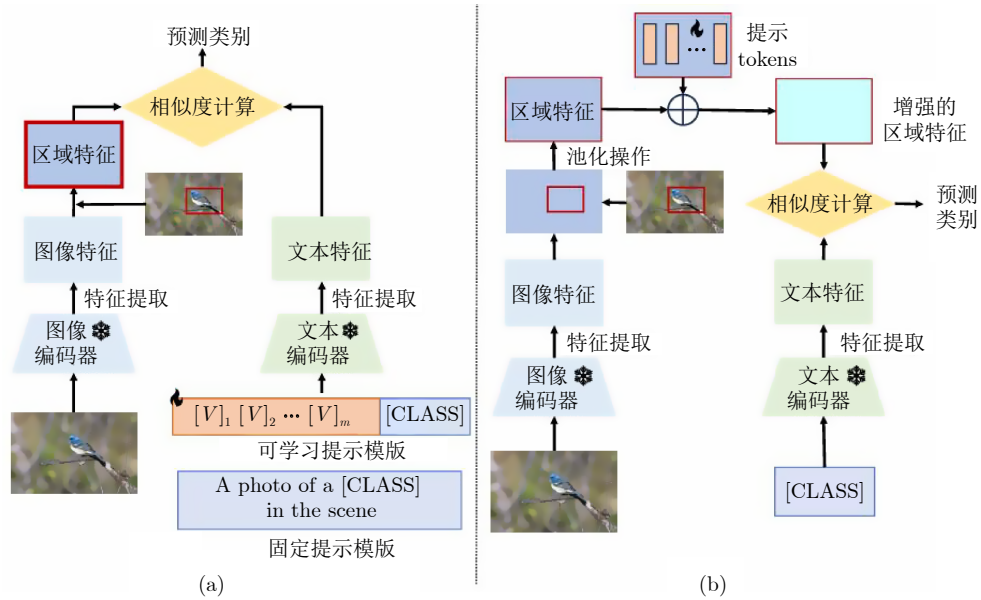

图 9 基于 CLIP 的 OVD 框架 ((a) 在 CLIP 的文本编码器端引入文本提示; (b) 在 CLIP 的图像编码器端引入提示 tokens)

Fig. 9 CLIP-based OVD framework ((a) Introducing text prompts at the text encoder side of CLIP; (b) Introducing prompt tokens at the image encoder side of CLIP)

配于各种视频理解任务: 视频动作识别、文本–视频检索和动作定位. 在资源受限的少样本和零样本场景下, 展现了显著的效果, 证明了 PL 在视频理解领域的潜力和实用性.

2.4.2 基于视觉–语言联合提示的视频动作识别方法

Wang 等[51] 利用两个预训练的单模态编码器充分发掘文本标签的语义信息, 通过手工设计的文本提示将此任务建模为多模态学习框架内的视频文本匹配问题. 值得注意的是, ActionCLIP 在文本编码器端引入手工设计的提示模板来扩展标签文本, 在视觉编码器端融入视觉提示, 使得预训练模型能够学习视频中所包含的关键时序信息. 通过应用两种模态的提示方法, ActionCLIP 在全监督的传统设置以及零样本和少样本设置中都展现出了出色的性能, 有效验证了 PL 方法在提高视频理解任务效率和性能中的关键作用.

2.5 图像描述任务

图像描述[52−55] (Image caption) 任务是指为图像生成一段描述性的文字, 这段文字应当能够表达图像中的主要事件、场景、对象和动作等. 针对图像描述任务的 PL 方法主要是基于文本提示的图像描述方法.

Mokady 等[52] 提出 CLIPCap 模型, 将基于视觉映射到语言空间的提示应用到图像描述任务中. 如图 10 所示, 该模型使用 CLIP 模型的图像编码器提取图像特征, 通过映射网络为每个图像生成一个前缀. 这些前缀与描述的嵌入向量进行拼接, 形成能被 LLM 理解和处理的前缀提示输入到语言模型中. 由于不同模型间有独立的潜在语义空间, 在训练映射网络的同时, 还需要微调语言模型. 在推理时, 语言模型从 CLIP 前缀开始逐个生成描述单词, 填入嵌入向量中. Tewel 等[53] 利用基于视觉映射到语言空间的提示, 在预训练的 CLIP 模型和 GPT-2 语言模型基础上, 提出 ZeroCap. 基于 CLIP 对输入图像的理解能力引导语言模型生成与给定图像相关的语言描述, 实现了零样本图像描述. 然而, 随着语言模型增大, 此类方法的计算成本会急剧增加. Su 等[54] 提出一个名为 MAGIC 的框架, 该框架可以使用图片模态的信息指导预训练语言模型完成一系列跨模态生成任务. 通过引入基于视觉映射到语

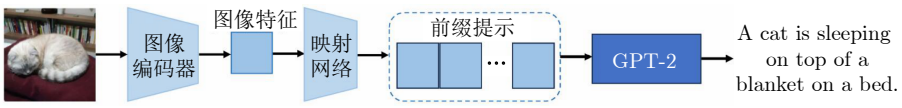

图 10 CLIPCap 图像描述任务框架

Fig. 10 Image caption task framework of CLIPCap

言空间的提示, 使用 CLIP 引导 LLM 模型生成与给定图像语义相关的结果. 同时, 提出一种新的解码方案, 将提示插入到解码过程中, 避免对上下文缓存和梯度进行更新, 因此与 ZeroCap 相比, MAGIC 拥有接近 27 倍的推理速度提升.

现有的方法在特定数据集上完成训练之后往往生成长度和风格固定的图像描述, 但迁移到其他数据集上时可能会失效. Wang 等[55] 提出在训练阶段优化多组连续提示来训练模型, 使其能在推理阶段通过使用不同的提示来生成符合特定长度和风格的视觉描述. 这种提示策略不仅提高了模型的适应性和灵活性, 而且允许模型在不同的风格和领域间进行无缝切换, 极大地提升了视频字幕生成的质量和用户体验.

2.6　多模态目标跟踪任务

多模态目标跟踪[56−59] 利用多种传感器或信息源获取的多模态数据来实现对目标位置的连续估计和追踪. 针对多模态目标跟踪任务的 PL 方法主要是基于视觉提示的多模态目标跟踪方法.

ProTrack[56] 方法引入基于提示模块的视觉提示, 将辅助模态的输入数据转换为主导的 RGB 模态, 从而将多模态跟踪任务转化为 RGB 单模态跟踪任务. 然而, ProTrack 提示参数由人为设定, 无法根据特定任务和数据自适应优化. 为了解决该问题, ViPT[57] (Visual prompt multi-modal tracking) 引入基于提示模块的视觉提示 MCP (Modality-complementary prompter), 将 RGB 模态与额外模态之间的关联转化为 tokens, 逐层添加到 Transformer 编码层的输入中, 如图 11 所示. TaTrack[58] 和 MPLT[59] 延续了 ViPT 中的提示模块设计, TaTrack 在多模态目标跟踪中引入时间信息; MPLT 在 ViPT 基础上引入双向模态交互设计, 能够获取更全面的双向模态融合表示, 可以进一步提高多模态目标跟踪性能.

2.7　视觉问答任务

视觉问答[60−62] 旨在基于图像回答开放式问题, 其目标是让模型接受视觉和文本问题的组合作为输入, 并输出用自然语言表达的答案. 针对视觉问答任务的 PL 方法主要是基于文本提示的视觉问答方法.

Tsimpoukelli 等[60] 利用基于视觉映射到语言空间的提示提出了 Frozen 模型. 通过从头训练一个视觉编码器将图像输入转换为语言空间的连续嵌入向量, 这些向量能够被冻结的 LLMs 直接理解和处理, 并结合文本信息生成对应的文本输出. 这种提示方法不仅在下游视觉任务中展现出了卓越的性

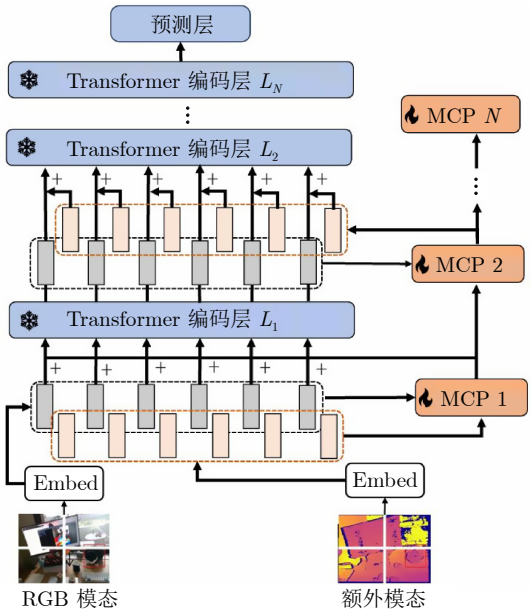

图 11　ViPT 方法流程图

Fig. 11　Flowchart of the ViPT method

能, 还为多模态研究提供了新的解决方案与启示. 与涉及分开的知识检索和答案推断的传统 VQA 方法不同, PICa[61] 将 GPT-3 视为一种隐含的、非结构化的知识库, 通过将图像转换为 GPT-3 可理解的文本提示并输入 GPT-3 以预测答案. PICa 继承了 GPT-3 的少样本学习能力, 在推断过程中仅使用少量上下文示例来适应 VQA 任务, 在包括 OK-VQA 在内的各种数据集上展现出强大的性能.

Jin 等[62] 提出基于手工设计的文本提示的 FEWVLM 方法, 与上述的 Frozen 和 PICa 方法相比, 旨在以更小的模型规模达到相似的准确性. 该方法采用编码器−解码器架构来编码视觉和文本输入, 并在文本提示的引导下结合图像内容生成准确的文本答复. 如图 12 所示, 为了获取更好的性能, FEWVLM 对比了多种基于手工设计的文本提示模板对输入文本进行处理, 以进一步指导模型生成更准确的答案. 一系列的对比实验结果证明, 当使用图中所示的 “Question: [Q] Answer: <text_1>”作为提示模板时, 模型对于视觉问答任务的准确率达到最高, 在保证参数量较小的情况下取得了与大模型 PICa 相近的性能.

2.8　视觉定位任务

视觉定位[63] (Visual grounding, VG) 是多模态学习的一个基本任务, 旨在将自然语言与图像区域对齐, 以识别图像中与给定描述对应的区域. 面向视觉定位的 PL 方法主要包括基于文本提示的视觉定位方法和基于视觉−语言联合提示的视觉定位方法.

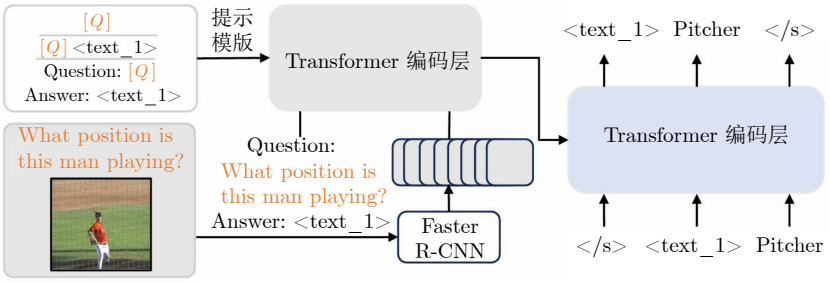

图 12 基于手工设计的文本提示的 FEWVLM 模型结构

Fig. 12 FEWVLM model structure based on hand-crafted text prompts

2.8.1 基于文本提示的视觉定位方法

Wang 等[63] 提出一种基于手工设计的文本提示方法 PTP (Position-guided text prompt), 旨在增强其识别图像中的物体位置及物体间空间关系的能力. PTP 首先将图像分为多块并识别每块中的物体, 然后通过一个文本提示模板“The block [number] has a [object]”将训练目标重新定义为一个填空问题. 此外, PTP 还设计了一个更复杂的文本提示模板“The block [number] has a [object 1] and a [object 2] in [top/bottom/left/right] of this block”, 用以捕获图片中不同物体之间的空间关系.

2.8.2 基于视觉–语言联合提示的视觉定位方法

Wu 等[64] 基于视觉–语言联合提示提出了 DMAP (Dynamic multi-modal prompting) 框架. 具体地, DMAP 设计一个动态提示网络, 通过基于输入实例从提示库中动态选择提示来生成多模态提示. 这些提示能够捕获文本与视觉输入之间的联系, 以增强自适应提示生成和多模态特征融合的性能, 有效提升了视觉定位任务的性能和训练效率.

2.9 3D 识别任务

3D 识别[65–66] 任务要求模型能够准确地理解三维场景中的空间和位置信息, 并准确识别出三维场景中的物体或场景. 面向 3D 识别任务的 PL 方法主要包括基于视觉提示的 3D 识别方法和基于视觉–语言联合提示的 3D 识别方法.

2.9.1 基于视觉提示的 3D 识别方法

Hegde 等[65] 在保持 CLIP 视觉和文本编码器冻结的情况下, 通过添加可学习的视觉提示 tokens 和 3D 编码器在训练过程中完成视觉、文本和 3D 数据的有效对齐. 在 CLIP 的视觉和文本编码器冻结的前提下, 成功地将模型扩展为一个 3D 识别网络, 名为 CG3D. 这种方法在零样本识别、语言场景理解和 3D 检索任务上均展现出卓越的性能. 此外, CG3D 训练过程中学习到的模型参数还能作为多种 3D 识别任务的初始权重, 为这些任务的发展和优化提供坚实的基础.

2.9.2 基于视觉–语言联合提示的 3D 识别方法

为了将大型预训练 CLIP 模型扩展到 3D 识别领域, Zhu 等[66] 巧妙地结合 CLIP 与 GPT-3, 提出 PointCLIP V2 模型. 此模型通过引入提示投影模块优化视觉深度图的生成, 同时利用 GPT-3 向文本描述中引入更多 3D 信息. 这一方法不仅显著增强了模型在 3D 识别上的性能, 还在零样本 3D 分割和 3D 目标检测等任务上展现了出色的迁移能力.

2.10 图像编辑任务

图像编辑[67] 是指根据文本输入对图像进行视觉和内容上的调整, 以增强其吸引力或传达特定信息, 包括去除瑕疵、调整构图、改变清晰度、调整色彩以及应用特效等. 针对图像编辑任务的 PL 方法主要是基于文本提示的图像编辑方法.

Text2LIVE[67] 通过内部数据集和预训练的 CLIP 模型训练生成器网络, 使其能够在文本提示的引导下自动定位并识别图像或视频中应编辑的对象或区域. 该网络生成包含颜色和透明度信息的 RGBA 编辑层, 然后通过合成操作将编辑层与原始输入结合, 实现对现实世界图像和视频的高保真、语义化和局部化编辑.

3 实验分析

本节通过实验比较和分析, 对图像分类、图像分割和多模态目标跟踪任务中典型的 PL 方法和非 PL 方法进行评估, 从定量的角度论证 PL 方法在降低模型的计算量、重用预训练模型以及提升模型的准确性和泛化能力等方面的优势.

3.1 图像分类任务

3.1.1 不同视觉提示方法对比

在图像分类中, 经典的非 PL 方法包括全面微

调 (Fully tuning) 和线性探测 (Linear probing), 经典的视觉 PL 方法包括 VP[21]、VPT[25] 及 DAM-VP[24].

方法简介. VP 在输入图像中添加像素扰动, VPT 在模型的输入 tokens 序列中添加少量的提示 tokens, DAM-VP 使用聚类方法将整个数据集划分为多个子集, 为每个子集分配并优化一个单独的提示. 全面微调方法更新预训练模型的所有参数, 线性探测方法只更新预训练模型的分类头部参数.

数据集. 本文选择 8 个常见的图像数据集来评估上述方法的性能, 包括通用图像识别数据集 CIFAR-10 和 CIFAR-100[68]、食物数据集 Food-101[69]、纹理数据集 DTD[70]、房屋门牌数字数据集 SVHN[71]、鸟类数据集 CUB-200[72]、狗类数据集 Stanford Dogs[73] 和花卉数据集 Flowers102[32].

评价指标. 使用 Top-1 准确率作为评价指标, 它表示分类器预测正确的样本占总样本数的比例, 用于衡量分类器在整个数据集上的整体性能.

执行细节. 本文选取在 ImageNet-22K 上预训练的 ViT-B/16 (ViT-B-22K) 和 Swin-Base (Swin-B-22K), 在训练时同时微调提示参数以及分类预测头部参数, 非 PL 方法训练 100 epochs, 而 PL 方法仅训练 50 epochs, 表 2 展示了不同方法的 Top-1 准确率对比结果.

结果分析. 早期的提示方法 VP 和 VPT 对于不同多样性的数据集使用固定数量的共享提示, 无法根据数据集特性动态调整提示数量, 在不同数据集上表现出性能的不稳定性. 例如, 在 CIFAR-100 数据集上, VP 显著优于非 PL 方法, 而在 Food-101 数据集上, 与非 PL 方法相比存在显著的性能差距; VPT 在 CIFAR-10 和 CIFAR-100 数据集上相对于线性探测方法的性能提升也存在较大差距. 而 DAM-VP 方法的提示设计充分考虑数据集多样性特性, 通过子集划分和单独优化子提示来动态分配提示数量, 能够以较少的迭代次数和微调参数量取得与全面微调相当甚至更优异的性能.

3.1.2　不同文本提示方法和视觉–语言联合提示方法对比

方法简介. 本文对比使用不同提示方法的四个经典的视觉–语言模型的图像分类性能, 分别是使用基于手工设计的文本提示的 CLIP[6] 模型、使用连续提示的 CoOp[13] 模型、基于图像引导的文本提示的 CoCoOp[14] 模型以及使用视觉–语言联合提示的 MaPLe[10] 模型.

数据集. 在 11 个典型的图像分类数据集上对模型进行评估, 这些数据集包括: 2 个通用对象数据集 ImageNet[12] 和 Caltech101[74]; 5 个细粒度数据集 OxfordPets[75]、StanfordCars[76]、Flowers102[32]、Food-101[69] 和 FGVCAircraft[77]; 1 个场景识别数据集 SUN397[78]; 1 个动作识别数据集 UCF101[79]; 1 个纹理数据集 DTD[70]; 1 个卫星图像数据集 EuroSAT[33].

评价指标. 以 Top-1 准确率为评价指标评估四个模型从基类到新类的泛化能力.

执行细节. 在实验中, 首先将数据集划分为基类 (Base) 和新类 (New), 由于 CLIP 使用基于手工设计的文本提示, 没有训练参数, 所以遵循零样本设置, CoOp、CoCoOp 和 MaPLe 则在基类上遵循少样本设置, 从每个类随机挑选 16 个样本进行训练, 而后在基类和新类上分别进行评估, 并计算准确率的均值.

结果分析. 表 3 直观地反映出在引入连续提示后, CoOp 在基类上的性能比 CLIP 有显著提升. 然而, 这种提升是以牺牲模型对新类的泛化能力为代

表 2　图像分类任务中 PL 方法和非 PL 示方法的性能对比 (加粗表示性能最优, 下划线表示性能次优) (%)
Table 2　In the task of image classification, a comparison of the performance between prompted and unprompted methods is presented (Bold indicates the best performance and underline indicates the second-best performance) (%)

	ViT-B-22K					Swin-B-22K				
	非 PL 方法		PL 方法			非 PL 方法		PL 方法		
	全面微调	线性探测	VP	VPT	DAM-VP	全面微调	线性探测	VP	VPT	DAM-VP
CIFAR-10	**97.4**	96.3	94.2	96.8	97.3	**98.3**	96.3	94.8	96.9	97.3
CIFAR-100	68.9	63.4	78.7	78.8	**88.1**	73.3	61.6	80.6	80.5	**88.1**
Food-101	84.9	84.4	80.5	83.3	**86.9**	**91.7**	88.2	83.4	90.1	90.5
DTD	64.3	63.2	59.5	65.8	**73.1**	72.4	73.6	75.1	78.5	**80.0**
SVHN	87.4	36.6	87.6	78.1	**87.9**	**91.2**	43.5	80.3	87.8	81.7
CUB-200	87.3	85.3	84.6	**88.5**	87.5	89.7	88.6	86.5	90.0	**90.4**
Stanford Dogs	89.4	86.2	84.5	90.2	**92.3**	86.2	85.9	81.3	84.8	**88.5**
Flowers102	98.8	97.9	97.7	99.0	**99.2**	98.3	99.4	98.6	99.3	**99.6**

表 3 从基类到新类的泛化设置下 CLIP、CoOp、CoCoOp 和 MaPLe 的对比 (HM 代表对基类和新类的准确率取调和平均值, 加粗表示性能最优) (%)

Table 3 Comparison of CLIP, CoOp, CoCoOp and MaPLe under the generalization setting from base class to new class (HM denotes the harmonic mean of the accuracies on both base and new classes, bold indicates the best performance) (%)

数据集	CLIP			CoOp			CoCoOp			MaPLe		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
ImageNet	72.43	68.14	70.22	76.47	67.88	71.92	75.98	70.43	73.10	**76.66**	**70.54**	**73.47**
Caltech101	96.84	94.00	95.40	**98.00**	89.81	93.73	97.96	93.81	95.84	97.74	**94.36**	**96.02**
OxfordPets	91.17	97.26	94.12	93.67	95.29	94.47	95.20	97.69	96.43	**95.43**	**97.76**	**96.58**
StanfordCars	63.37	**74.89**	68.65	**78.12**	60.40	68.13	70.49	73.59	72.01	72.94	74.00	**73.47**
Flowers102	72.08	**77.80**	74.83	**97.60**	59.67	74.06	94.87	71.75	81.71	95.92	72.46	**82.56**
Food-101	90.10	91.22	90.66	88.33	82.26	85.19	90.70	91.29	90.99	**90.71**	**92.05**	**91.38**
FGVCAircraft	27.19	**36.29**	31.09	**40.44**	22.30	28.75	33.41	23.71	27.74	37.44	35.61	**36.50**
SUN397	69.36	75.35	72.23	80.60	65.89	72.51	79.74	76.86	78.27	**80.82**	**78.70**	**79.75**
DTD	53.24	**59.90**	56.37	79.44	41.18	54.24	77.01	56.00	64.85	**80.36**	59.18	**68.16**
EuroSAT	56.48	64.05	60.03	92.19	54.74	68.69	87.49	60.04	71.21	**94.07**	**73.23**	**82.35**
UCF101	70.53	77.50	73.85	**84.69**	56.05	67.46	82.33	73.45	77.64	83.00	**78.66**	**80.77**
平均值	69.34	74.22	71.10	**82.69**	63.22	71.66	80.47	71.69	75.83	82.28	**75.14**	**78.55**

价的. CoCoOp 通过对连续提示的改进, 使用基于图像引导的文本提示来弥补连续提示的不足. 在基类上保证与 CoOp 相近性能的同时, 进一步提升了模型在新类上的泛化能力, 接近甚至达到了零样本的 CLIP 模型的水平. 而后, MaPLe 通过使用视觉–语言联合提示, 进一步提升了模型在基类和新类上的表现, 表明这种同时引入两种模态提示的方法对于提升视觉–语言模型的泛化性能的有效性, 为进一步探索将多模态 PL 应用于 CV 任务提供了坚实的基础和动力.

3.2 图像分割任务

图像分割大致分为语义分割、实例分割和全景分割三大子任务, 本节以语义分割和实例分割任务为例, 探讨代表性的 PL 方法和非 PL 方法将预训练模型迁移到下游任务时的性能表现.

方法简介. SPM[35] 和 AdaptFormer[11] 在预训练的主干网络中引入轻量级的提示模块. VPT 在主干网络中添加提示 tokens. fully tuning 更新模型的全部参数. head tuning 只更新模型的分割头部参数. SAM[7] 和 EfficientSAM[41] 使用点/框/文本或掩码作为提示, HQ-SAM[42] 在 SAM 的掩码解码器中引入提示 tokens, PA-SAM[43] 在 SAM 的掩码解码器旁引入提示模块. Mask2Former[80] 和 OneFormer[81] 为通用图像分割模型.

数据集. 对于语义分割, 选取经典的场景解析数据集 ADE20K[82], 包括 150 个语义类别. 对于实例分割, 选取经典的 COCO[83] 数据集, 包括 80 个类别.

评价指标. 对于 ADE20K 和 COCO 数据集, 分别用均交并比 (mIoU) 和 mAP 作为评价指标.

执行细节. 语义分割中的 SPM、AdaptFormer、VPT、fully tuning 和 head tuning 均基于预训练的 ViT-L. 实例分割中的 SAM、EfficientSAM、HQ-SAM、PA-SAM 使用现有的检测模型产生的检测框作为固定提示框. 此外, HQ-SAM 引入可学习的提示 tokens, PA-SAM 引入可学习的提示模块, 非 PL 方法 Mask2Former 和 OneFormer 使用 Swin-L 作为主干网络.

结果分析. 在语义分割任务中, 表 4 直观地反映出所有的 PL 方法要优于非 PL 方法 head tuning, SPM、VPT 和 AdaptFormer 与 fully tuning 相比仍存在一定程度的性能劣势. 但是, 这些方法可以显著减少参数优化带来的计算开销. SAM 系列模型可以超出以往最优的 fully tuning. 引入提示模块的 SPM 和 AdaptFormer 要优于引入提示 tokens 的 VPT, 表明添加到预训练模型中的提示模块能够产生更有效的提示, 进而生成更准确有效的特征, 引导预训练模型产生决策结果. 在实例分割任务中, 表 5 直观地反映出引入提示的 SAM、HQ-SAM 和 PA-SAM 性能优于非 PL 方法 Mask2Former 和 OneFormer. 然而 SAM 仅仅基于固定框提示产生分割结果, HQ-SAM 和 PA-SAM 则结合固定框提示和可学习的提示产生分割结果, 性能高于 SAM, 表明可学习的提示优于固定提示. 另外, 引入提示模块的 PA-SAM 性能优于引入提示 tokens

表 4　ADE20K 数据集上 PL 方法和非 PL 方法的语义分割性能对比

Table 4　Comparison of semantic segmentation performance on the ADE20K dataset between prompted and unprompted methods

		参数量 (M)	mIoU (%)
PL 方法	SPM	14.90	45.05
	VPT	13.39	42.11
	AdaptFormer	16.31	44.00
	SAM	—	**53.00**
	EfficientSAM	—	51.80
非 PL 方法	fully tuning	317.29	47.53
	head tuning	13.14	37.77

表 5　COCO 数据集上 PL 方法和非 PL 方法的实例分割性能对比

Table 5　Comparison of instance segmentation performance on the COCO dataset between prompted and unprompted methods

		mAP (%)
PL 方法	SAM	46.8
	EfficientSAM	44.4
	HQ-SAM	49.5
	PA-SAM	**49.9**
非 PL 方法	Mask2Former	43.7
	OneFormer	45.6

的 HQ-SAM, 反映出提示模块相较于提示 tokens 的优势.

3.3　多模态目标跟踪任务

RGB-Thermal[59] 多模态目标跟踪利用可见光图像和热红外图像的互补性, 综合考虑两种模态的信息获取更准确的目标定位, 适用于光照变化和目标遮挡等复杂环境.

方法简介. ProTrack[56] 和 ViPT[57] 均使用冻结的预训练的 RGB 跟踪模型作为主干网络. ProTrack 方法使用提示函数将额外模态数据转换为 RGB 模态. ViPT 方法使用可学习的 MCP 模块逐阶段获取 RGB 模态和额外模态的互补表示, 产生提示 tokens, 并添加到预训练 RGB 跟踪模型的输入 tokens 序列.

数据集. RGBT234[84] 和 LasHeR[85] 包含不同场景和环境条件下的 RGB 图像和热红外图像的对应帧对, 包括室内、室外、光照变化和天气变化等.

评价指标. 使用准确率 (precision) 和成功率 (success) 作为评价指标. 准确率衡量跟踪器在目标位置估计上的精确程度. 成功率衡量跟踪器成功跟踪目标的能力, 即在一定重叠率阈值下, 跟踪器成功跟踪的帧数占总帧数的比例.

执行细节. PL 方法使用与 OsTrack 相同的模型架构和损失函数, 此外还选用了代表性的 OsTrack、FANet 和 SGT 作为 RGB-T 跟踪的非 PL 学习方法, 跟踪性能对比结果如表 6 所示.

表 6　多模态跟踪任务中 PL 方法和非 PL 方法的性能对比 (%)

Table 6　Performance comparison between prompted and unprompted methods in multimodal tracking tasks (%)

		RGBT234		LasHeR	
		precision	success	precision	success
PL 方法	TaTrack	87.2	64.4	**85.3**	**61.8**
	MPLT	**88.4**	**65.7**	72.0	57.1
	ViPT	83.5	61.7	65.1	52.5
	ProTrack	79.5	59.9	53.8	42.0
非 PL 方法	OsTrack	72.9	54.9	51.5	41.2
	FANet	78.7	55.3	44.1	30.9
	SGT	72.0	47.2	36.5	25.1

结果分析. 表 6 直观地反映出现有的多模态目标跟踪 PL 方法在 RGB-T 任务上的跟踪性能显著优于已有的非 PL 方法. 以 OsTrack 方法为基线, RGBT234 数据集上最优的 MPLT 获取了 15.5% 的准确率提升和 10.8% 的成功率提升, LasHeR 数据集上最优的 TaTrack 获取了 33.8% 的准确率提升和 20.6% 的成功率提升. 此外, 不同类型的提示存在显著的性能差异, 可训练的提示模块优于人工设计的固定提示模块. ProTrack 和 MPLT 基于可优化的提示模块, 可以有效捕捉模态互补信息. ProTrack 方法基于固定的提示模块将不同模态信息融合为单一模态, 无法获取有效的模态融合表示, 因而效果较差.

4　当前问题及未来研究方向

综上所述, PL 作为一种新兴的学习范式, 在 CV 领域受到广泛关注, 但其工作机制和潜在挑战仍待深入研究, 同时也为未来研究指明方向. 本节将详细探讨视觉领域提示机制存在的问题, 并对其未来发展方向提出一些见解.

4.1　挑战

1) 通用多任务 PL 的设计与学习难度大. 首先, 由于不同的 CV 任务在训练数据、输入输出形式等方面存在明显差异, 导致现有的多模态大模型和

PL 方法都是针对特定的下游任务进行设计和优化的, 因此难以有效地迁移到其他类别的任务中复用. 其次, 针对特定任务设计的 PL 方法本身还在一定程度上受到模型网络结构的制约, 如文本提示只能用于包含文本处理能力的多模态模型中, 在单模态的纯视觉模型中效果欠佳; 基于提示 tokens 的视觉提示高度依赖于 Transformer 结构, 难以迁移到其他类型的网络结构中. 这些问题都进一步限制了当前 PL 方法的通用性. 因此, 如何设计通用的 PL 方法, 使其能够有效地泛化到不同的 CV 应用, 是当前面临的一大挑战.

2) 面向 CV 领域的提示设计缺乏可解释性. 为了适应 CV 领域的不同下游任务, 现有提示学习设计通常致力于引导模型提取与预训练模型知识表示匹配的特征. 然而, 由于当前效果较好的可学习提示大多是以黑盒方式进行优化的, 如连续提示、基于提示 tokens 的视觉提示等, 很少有工作探索优化过程中特征的演变和最终优化得到的下游任务特征与预训练任务特征的匹配程度, 因而缺乏可解释性. 这种缺陷可能会引发一系列问题, 包括隐私数据的泄露风险、误导性的"幻觉"和对抗攻击干扰等, 从而大大降低模型的可靠性. 可见, 现有提示方法缺乏可解释性是其面临的又一重要挑战.

3) 多种跨模态 PL 方法交互与融合难度高. 现有研究表明, 与单独使用视觉或文本提示相比, 将文本和视觉端的 PL 方法同时引入, 并设计有效的跨模态交互与融合机制会显著提升 PL 方法的性能. 随着 CV 研究的深入, 新兴模态数据如音频和生理信号等正逐渐融入多种复杂 CV 任务中. 因此, 针对复杂场景和任务的跨模态提示设计也越来越多. 面对不断增长的应用场景需求, 如何有效地将多种单模态 PL 方法进行高效的交互与融合, 以实现跨模态提示特征的精确对齐, 是 PL 发展的必然挑战.

4) PL 在生成式视觉任务中的探索有限. 当前, PL 在 CV 中的应用集中在判别式任务上, 如图像分类和图文匹配, 并取得了不错的进展. 然而, 对于更具挑战性的生成式任务, 如图像生成、图像编辑等, PL 的探索和应用还相对有限. 当前对于 PL 的研究依旧集中于使用简单的文本提示或借助于 LLMs 对语言的理解能力, 来引导生成式大模型更好地理解用户的指令, 鲜有工作尝试将可学习的提示深度地引入图像的修改与生成过程中. 因此, 如何设计和创新提示方法, 使其能够指导生成式模型更好地处理和生成更复杂的视觉内容是当前的一大挑战.

4.2 未来展望

1) 引入多任务学习机制和持续学习来引导多任务 PL 训练. 尽管 MVLPT[20] 方法使用基于多任务的文本提示已经取得了一定进展. 然而, 该方法主要通过相似的任务共享一组提示参数来实现多任务学习, 对任务的输入输出形式有一定的限制, 对新任务的扩展能力不足. 未来, 可以探索更加通用的预训练大模型, 引入多任务学习机制和持续学习来引导实现不同任务 PL 的协同, 实现更广泛的多任务提示训练.

2) 探索人类思维链 (Chain of thought, CoT) 和自监督的提示优化方法, 提升 PL 的可解释性. CoT 通过在模型生成过程中显式地记录推理步骤, 使得模型不仅给出答案, 还展示出解决问题的思维过程, 可以有效提升 PL 设计的可解释性. 此外, 还可以利用自监督学习方法, 通过生成解释性提示和输出对, 训练模型生成更加可解释的结果.

3) 研究高效的跨模态 PL 交互和耦合学习方式. 随着 CV 领域研究的不断深入和新模态数据的大量涌现, 当前研究的重点将转向开发多模态的 PL 方法, 促进这些方法之间的高效交互与融合是未来 PL 研究的重要方向. 如, 引入交叉注意力机制对不同模态提示表征进行注意力计算, 提取模态间的关联信息, 从而促进跨模态提示信息的交互. 此外, 还可以通过设计更高效的模型结构和优化算法, 在降低资源消耗的同时提升模型性能.

4) 探索生成式学习机制与提示学习结合在生成式视觉任务中的应用和创新. 生成式学习机制是指利用生成模型对数据进行建模和生成的学习方法, 如扩散学习、对抗学习等. 未来, 可以引入生成式学习机制协助 PL 学习数据的分布, 提升模型对不同下游任务的理解和推理能力, 从而指导模型实现不同的生成任务. 另外, 可以进一步挖掘生成式学习机制在 PL 的潜力, 设计新颖的生成式提示方法, 帮助精准捕捉用户意图并增强模型的创造力.

5 总结

传统的"预训练+微调"范式存在对计算资源和标签数据需求高、计算难度大的问题, 尤其是随着模型规模和复杂度的提高, 这些问题变得更加严峻. PL 作为一种能有效替代微调的方法, 对于提升模型性能和泛化能力、重用预训练模型以及降低计算量起到了关键作用. 因此, 本文首先对 PL 的发展历程进行梳理, 并总结分析在 CV 领域中不同类型提示方法的原理和优缺点. 随后, 针对不同的视觉任务, 按照提示类型对经典的 PL 方法进行总结. 在实

验部分, 进一步对代表性的 PL 方法和非 PL 方法进行实验对比和分析, 验证 PL 方法在提升模型准确率和泛化性方面的优势. 最后, 基于上述分析, 总结现有 PL 方法在 CV 领域存在的挑战, 并对 PL 的未来发展方向提出一些见解. 相信本文能够帮助研究人员全面深入地了解 PL 方法在 CV 任务中的工作机制, 为推动 PL 方法在 CV 领域的进一步发展提供参考.

References

1 Xu M W, Yin W S, Cai D Q, Yi R J, Xu D L, Wang Q P, et al. A survey of resource-efficient LLM and multimodal foundation models. arXiv preprint arXiv: 2401.08092, 2024.

2 Zhou J H, Chen Y Y, Hong Z C, Chen W H, Yu Y, Zhang T, et al. Training and serving system of foundation models: A comprehensive survey. *IEEE Open Journal of the Computer Society*, 2024, **5**: 107−119

3 Liu Z M, Yu X T, Fang Y, Zhang X M. GraphPrompt: Unifying pre-training and downstream tasks for graph neural networks. In: Proceedings of the ACM Web Conference. Austin, USA: ACM, 2023. 417−428

4 Liu P F, Yuan W Z, Fu J L, Jiang Z B, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2023, **55**(9): Article No. 195

5 Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv: 2304.07193, 2023.

6 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748−8763

7 Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3992−4003

8 Liao Ning, Cao Min, Yan Jun-Chi. Visual prompt learning: A survey. *Chinese Journal of Computers*, 2024, **47**(4): 790−820
(廖宁, 曹敏, 严骏驰. 视觉提示学习综述. 计算机学报, 2024, **47**(4): 790−820)

9 Zang Y H, Li W, Zhou K Y, Huang C, Loy C C. Unified vision and language prompt learning. arXiv preprint arXiv: 2210.07225, 2022.

10 Khattak M U, Rasheed H, Maaz M, Khan S, Khan F S. MaPLe: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 19113−19122

11 Chen S F, Ge C J, Tong Z, Wang J L, Song Y B, Wang J, et al. AdaptFormer: Adapting vision Transformers for scalable visual recognition. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1212

12 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248−255

13 Zhou K Y, Yang J K, Loy C C, Liu Z W. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022, **130**(9): 2337−2348

14 Zhou K Y, Yang J K, Loy C C, Liu Z W. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16795−16804

15 Derakhshani M M, Sanchez E, Bulat A, da Costa V G, Snoek C G M, Tzimiropoulos G, et al. Bayesian prompt learning for image-language model generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 15191−15200

16 Yao H T, Zhang R, Xu C S. Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 6757−6767

17 Bulat A, Tzimiropoulos G. LASP: Text-to-text optimization for language-aware soft prompting of vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 23232−23241

18 Zhu B E, Niu Y L, Han Y C, Wu Y, Zhang H W. Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 15613−15623

19 Huang T, Chu J, Wei F Y. Unsupervised prompt learning for vision-language models. arXiv preprint arXiv: 2204.03649, 2022.

20 Shen S, Yang S J, Zhang T J, Zhai B H, Gonzalez J E, Keutzer K, et al. Multitask vision-language prompt tuning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 5644−5655

21 Bahng H, Jahanian A, Sankaranarayanan S, Isola P. Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv: 2203.17274, 2022.

22 Chen A C, Yao Y G, Chen P Y, Zhang Y H, Liu S J. Understanding and improving visual prompting: A label-mapping perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 19133−19143

23 Oh C, Hwang H, Lee H Y, Lim Y, Jung G, Jung J, et al. BlackVIP: Black-box visual prompting for robust transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 24224−24235

24 Huang Q D, Dong X Y, Chen D D, Zhang W M, Wang F F, Hua G, et al. Diversity-aware meta visual prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 10878−10887

25 Jia M L, Tang L M, Chen B C, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 709−727

26 Tu C H, Mai Z D, Chao W L. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 7725−7735

27 Das R, Dukler Y, Ravichandran A, Swaminathan A. Learning expressive prompting with residuals for vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 3366−3377

28 Dong B W, Zhou P, Yan S C, Zuo W M. LPT: Long-tailed prompt tuning for image classification. In: Proceedings of the Eleventh International Conference on Learning Representations.

Kigali, Rwanda: ICLR, 2023. 1−20

29 Zhang Y H, Zhou K Y, Liu Z W. Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2024.3435939

30 Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685, 2021.

31 Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2790−2799

32 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India: IEEE, 2008. 722−729

33 Helber P, Bischke B, Dengel A, Borth D. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, **12**(7): 2217−2226

34 Fahes M, Vu T H, Bursuc A, Pérez P, de Charette R. PØDA: Prompt-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18577−18587

35 Liu L B, Chang J L, Yu B X B, Lin L, Tian Q, Chen C W. Prompt-matched semantic segmentation. arXiv preprint arXiv: 2208.10159, 2022.

36 Liu W H, Shen X, Pun C M, Cun X D. Explicit visual prompting for low-level structure segmentations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 19434−19445

37 Bar A, Gandelsman Y, Darrell T, Globerson A, Efros A A. Visual prompting via image inpainting. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1813

38 Ma X H, Wang Y M, Liu H, Guo T Y, Wang Y H. When visual prompt tuning meets source-free domain adaptive semantic segmentation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 293

39 Zhao X, Ding W C, An Y Q, Du Y L, Yu T, Li M, et al. Fast segment anything. arXiv preprint arXiv: 2306.12156, 2023.

40 Zhang C N, Han D S, Qiao Y, Kim J U, Bae S H, Lee S, et al. Faster segment anything: Towards lightweight SAM for mobile applications. arXiv preprint arXiv: 2306.14289, 2023.

41 Xiong Y Y, Varadarajan B, Wu L M, Xiang X Y, Xiao F Y, Zhu C C, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16111−16121

42 Ke L, Ye M Q, Danelljan M, Liu Y F, Tai Y W, Tang C K, et al. Segment anything in high quality. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1303

43 Xie Z Z, Guan B C, Jiang W H, Yi M Y, Ding Y, Lu H T, et al. PA-SAM: Prompt adapter SAM for high-quality image segmentation. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE, 2024. 1−6

44 Wang X L, Zhang X S, Cao Y, Wang W, Shen C H, Huang T J. SegGPT: Segmenting everything in context. arXiv preprint arXiv: 2304.03284, 2023.

45 Ren T H, Liu S L, Zeng A L, Lin J, Li K C, Cao H, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv: 2401.14159, 2024.

46 Zou X Y, Yang J W, Zhang H, Li F, Li L J, Wang J F, et al. Segment everything everywhere all at once. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 868

47 Gu X Y, Lin T Y, Kuo W C, Cui Y. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv: 2104.13921, 2021.

48 Du Y, Wei F Y, Zhang Z H, Shi M J, Gao Y, Li G Q. Learning to prompt for open-vocabulary object detection with vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 14064−14073

49 Wu X S, Zhu F, Zhao R, Li H S. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 7031−7040

50 Ju C, Han T D, Zheng K H, Zhang Y, Xie W D. Prompting visual-language models for efficient video understanding. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 105−124

51 Wang M M, Xing J Z, Liu Y. ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv: 2109.08472, 2021.

52 Mokady R, Hertz A, Bermano A H. ClipCap: CLIP prefix for image captioning. arXiv preprint arXiv: 2111.09734, 2021.

53 Tewel Y, Shalev Y, Schwartz I, Wolf L. ZeroCap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17897−17907

54 Su Y X, Lan T, Liu Y H, Liu F Y, Yogatama D, Wang Y, et al. Language models can see: Plugging visual controls in text generation. arXiv preprint arXiv: 2205.02655, 2022.

55 Wang N, Xie J H, Wu J H, Jia M B, Li L L. Controllable image captioning via prompting. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 2617−2625

56 Yang J Y, Li Z, Zheng F, Leonardis A, Song J K. Prompting for multi-modal tracking. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: Association for Computing Machinery, 2022. 3492−3500

57 Zhu J W, Lai S M, Chen X, Wang D, Lu H C. Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 9516−9526

58 He K J, Zhang C L, Xie S, Li Z X, Wang Z W. Target-aware tracking with long-term context attention. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 773−780

59 Luo Y, Guo X Q, Feng H, Ao L. RGB-T tracking via multi-modal mutual prompt learning. arXiv preprint arXiv: 2308.16386, 2023.

60 Tsimpoukelli M, Menick J, Cabi S, Eslami S M A, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. arXiv preprint arXiv: 2106.13884, 2021.

61 Yang Z Y, Gan Z, Wang J F, Hu X W, Lu Y M, Liu Z C, et al. An empirical study of GPT-3 for few-shot knowledge-based VQA. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 3081−3089

62 Jin W, Cheng Y, Shen Y L, Chen W Z, Ren X. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: ACL, 2022. 2763–2775

63 Wang A J, Zhou P, Shou M Z, Yan S C. Enhancing visual grounding in vision-language pre-training with position-guided text prompts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(5): 3406–3421

64 Wu W S, Liu T, Wang Y K, Xu K, Yin Q J, Hu Y. Dynamic multi-modal prompting for efficient visual grounding. In: Proceedings of the 6th Chinese Conference on Pattern Recognition and Computer Vision. Xiamen, China: Springer, 2023. 359–371

65 Hegde D, Valanarasu J M J, Patel V M. CLIP goes 3D: Leveraging prompt tuning for language grounded 3D recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023. 2020–2030

66 Zhu X Y, Zhang R R, He B W, Guo Z Y, Zeng Z Y, Qin Z P, et al. PointCLIP V2: Prompting clip and GPT for powerful 3D open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 2639–2650

67 Bar-Tal O, Ofri-Amar D, Fridman R, Kasten Y, Dekel T. Text2LIVE: Text-driven layered image and video editing. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 707–723

68 Krizhevsky A. Learning Multiple Layers of Features From Tiny Images, Technical Report TR-2009, University of Toronto, Canada, 2009.

69 Bossard L, Guillaumin M, van Gool L. Food-101: Mining discriminative components with random forests. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 446–461

70 Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 3606–3613

71 Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng A Y. Reading digits in natural images with unsupervised feature learning. In: Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain: NIPS, 2011. Article No. 4

72 Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2010-001, California Institute of Technology, USA, 2010.

73 Khosla A, Jayadevaprakash N, Yao B P, Li F F. Novel dataset for fine-grained image categorization. In: Proceedings of the First Workshop on Fine-grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE, 2011.

74 Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop. Washington, USA: IEEE, 2004. 178

75 Parkhi O M, Vedaldi A, Zisserman A, Jawahar C V. Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 3498–3505

76 Krause J, Stark M, Deng J, Li F F. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013. 554–561

77 Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. arXiv preprint arXiv: 1306.5151, 2013.

78 Xiao J X, Hays J, Ehinger K A, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3485–3492

79 Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv: 1212.0402, 2012.

80 Cheng B W, Misra I, Schwing A G, Kirillov A, Girdhar R. Masked-attention mask Transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1280–1289

81 Jain J, Li J C, Chiu M, Hassani A, Orlov N, Shi H. OneFormer: One Transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 2989–2998

82 Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 5122–5130

83 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 740–755

84 Xiao Y, Yang M M, Li C L, Liu L, Tang J. Attribute-based progressive fusion network for RGBT tracking. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 2831–2838

85 Li C L, Xue W L, Jia Y Q, Qu Z C, Luo B, Tang J, et al. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 2022, **31**: 392–404

刘袁缘 中国地质大学 (武汉) 计算机学院副教授. 主要研究方向为计算机视觉. E-mail: liuyy@cug.edu.cn

(**LIU Yuan-Yuan** Associate professor at the School of Computer Science, China University of Geosciences. Her main research interest is computer vision.)

刘树阳 中国地质大学 (武汉) 计算机学院硕士研究生. 主要研究方向为人脸情感识别.
E-mail: 20171003670@cug.edu.cn

(**LIU Shu-Yang** Master student at the School of Computer Science, China University of Geosciences. His main research interest is facial emotion recognition.)

刘云娇 中国地质大学 (武汉) 计算机学院硕士研究生. 主要研究方向为遥感图像分割.
E-mail: luyunjiao@cug.edu.cn
(**LIU Yun-Jiao** Master student at the School of Computer Science, China University of Geosciences. Her main research interest is remote sensing image segmentation.)

袁雨晨 中国地质大学 (武汉) 计算机学院硕士研究生. 主要研究方向为聚类分析.
E-mail: 1202321648@cug.edu.cn
(**YUAN Yu-Chen** Master student at the School of Computer Science, China University of Geosciences. His main research interest is cluster analysis.)

唐 厂 中国地质大学 (武汉) 计算机学院教授. 主要研究方向为多视图学习. E-mail: tangchang@cug.edu.cn
(**TANG Chang** Professor at the School of Computer Science, China University of Geosciences. His main research interest is multi-view learning.)

罗 威 中国舰船研究设计中心高级工程师. 主要研究方向为舰船人工智能. 本文通信作者.
E-mail: csddc_weiluo@163.com
(**LUO Wei** Senior engineer at China Ship Development and Design Center. His main research interest is ship artificial intelligence. Corresponding author of this paper.)