北京航空航天大學
BEIHANG UNIVERSITY

Beihang University

开放视觉感知

北京航空航天大学
刘偲

大纲

① 任务介绍

② 研究现状

③ 总结与展望

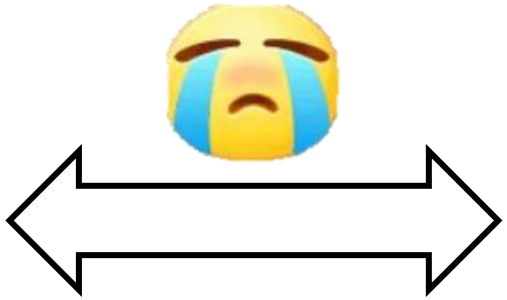
01

任务介绍

任务介绍

传统机器感知过程过于**依赖数据**，概念空间封闭，无法完成**陌生概念**的正确识别

机器感知过程

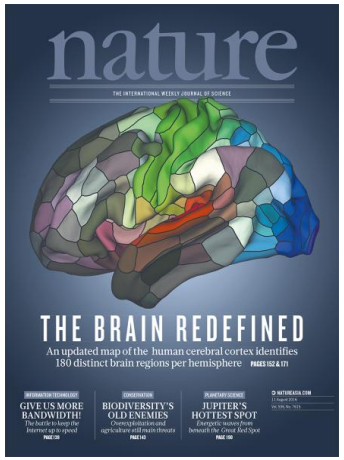


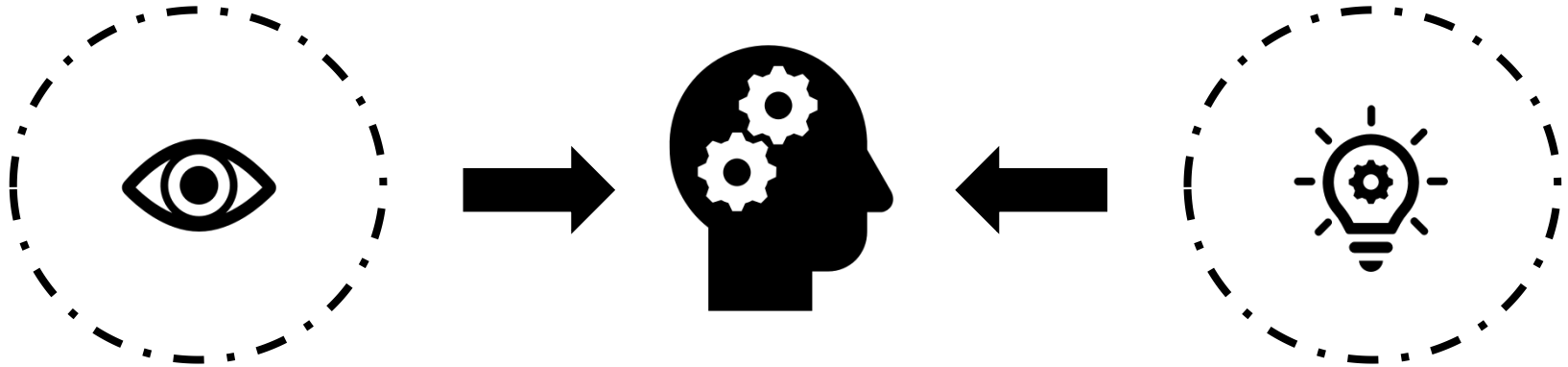
封闭概念空间

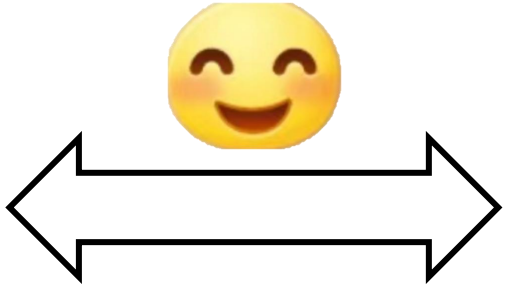
开放视觉感知

人类将环境感受与知识信息融合，并理解新概念

B. T. Thomas Yeo







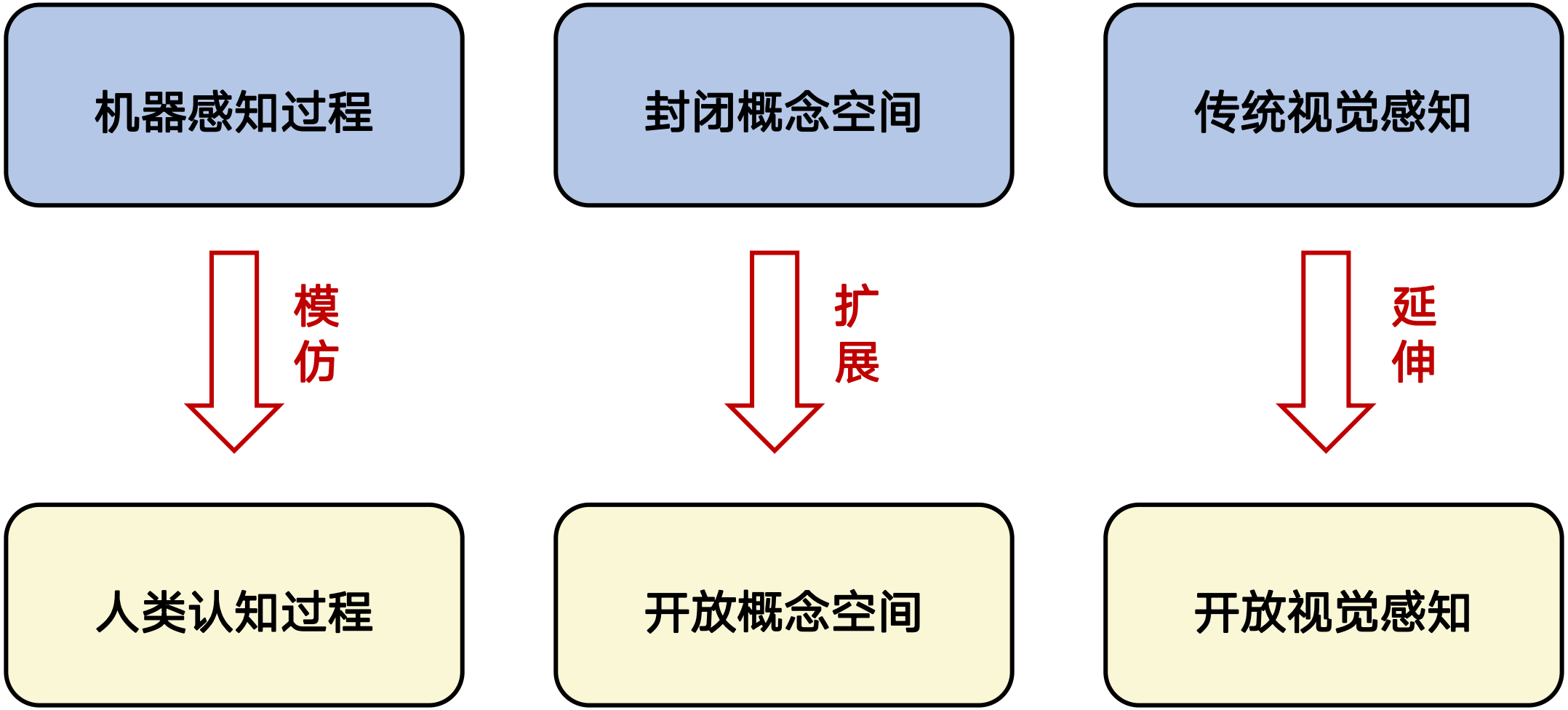
人类认知过程

开放概念空间

Yeo B T T, Eickhoff S B. A modern map of the human cerebral cortex[J]. Nature, 2016, 536(7615): 152-154.

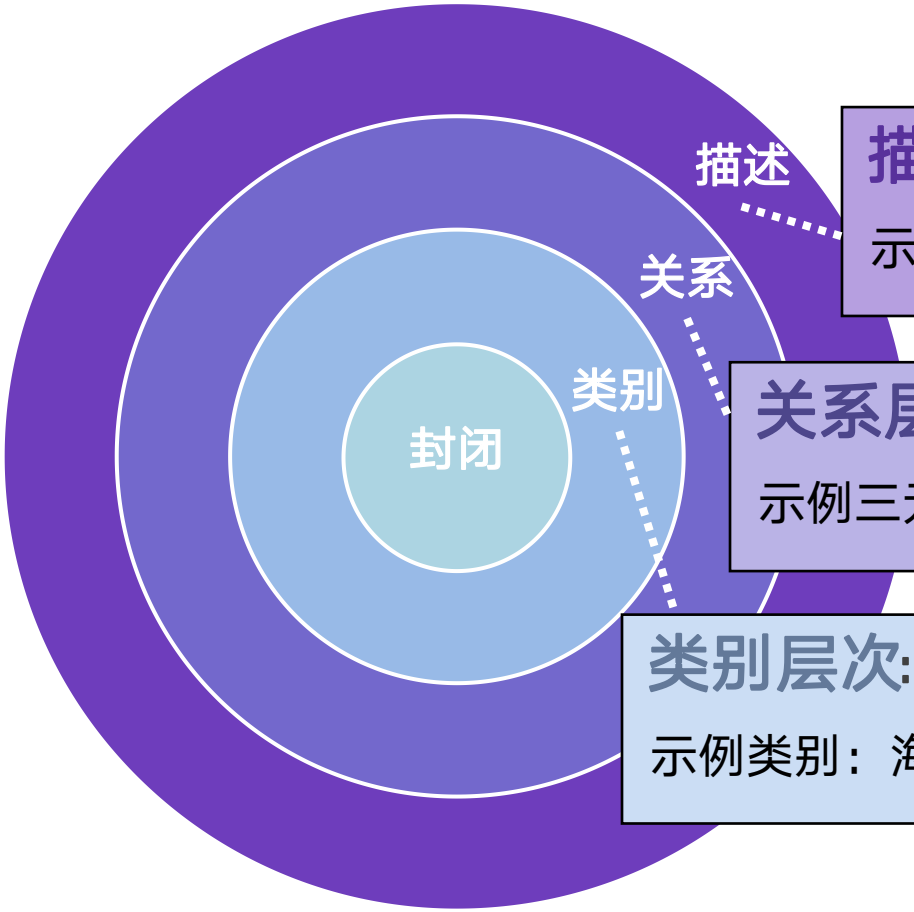
开放视觉感知

机器感知过程与人类认知过程的对比



开放视觉感知

开放视觉感知是概念空间的开放



描述层次 任意描述

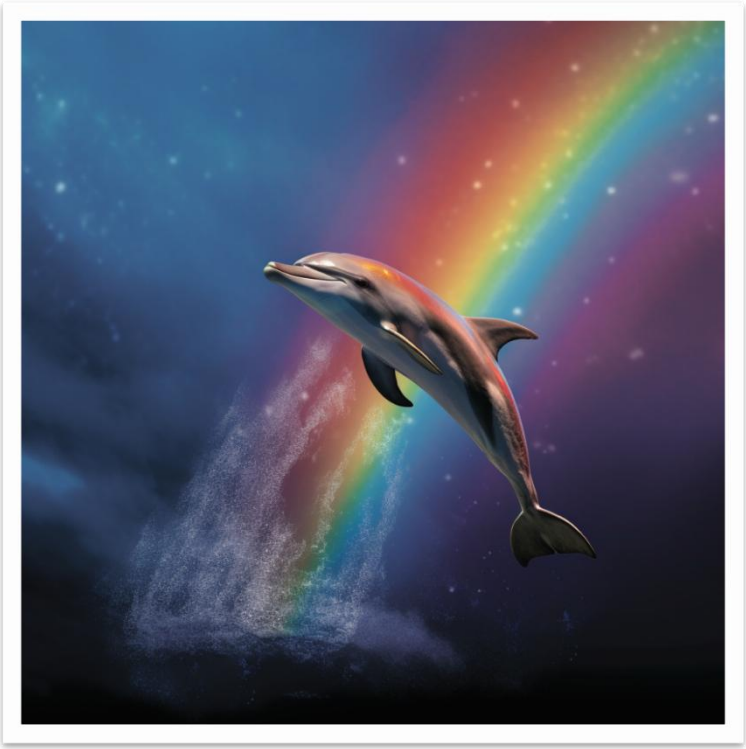
示例描述：“在飞跃彩虹的海豚”

关系层次: 任意三元组

示例三元组：<海豚，跃过，彩虹>

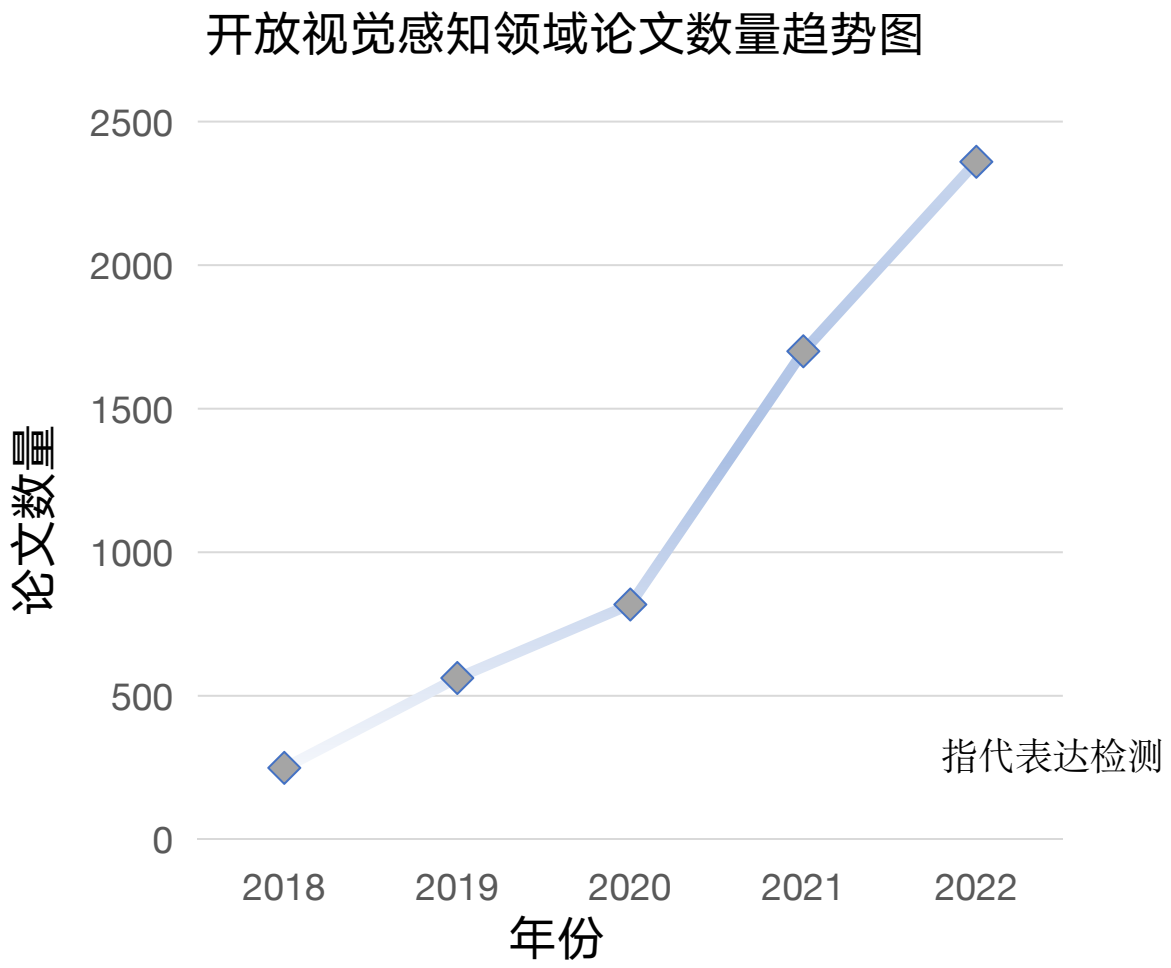
类别层次: 任意类别

示例类别：海豚，彩虹



开放视觉感知

开放视觉感知收到广泛关注，涌现了众多的研究方向和工作



论文数量数据来源于Web of Science

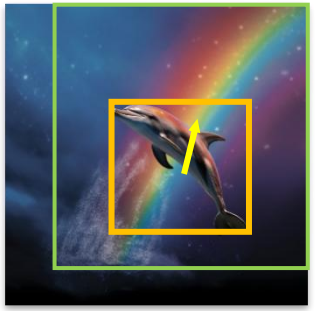
类别层次示例

开放词表分割 ...

海豚、彩虹

关系层次示例

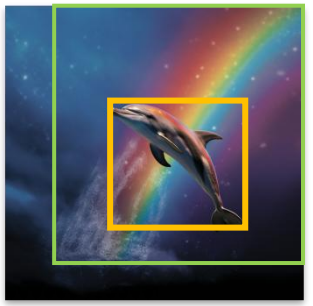
开放关系检测 ...



<海豚，跃过，彩虹>

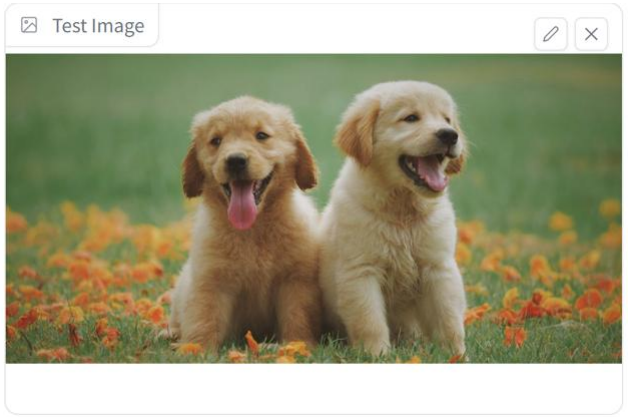
描述层次示例

指代表达检测 ...

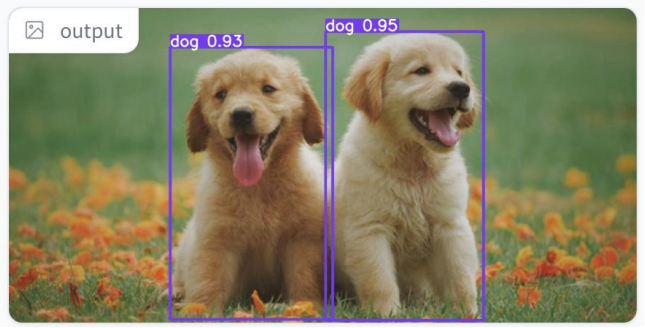


“正在飞跃彩虹的海豚”

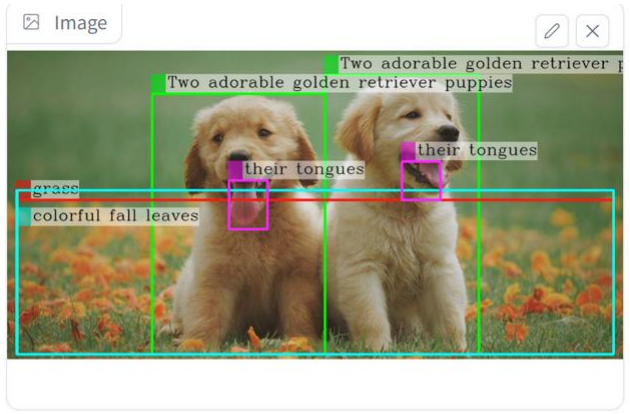
开放视觉感知



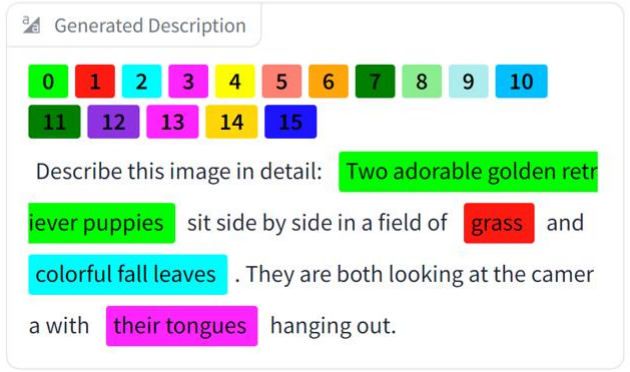
传统方法



开放视觉感知



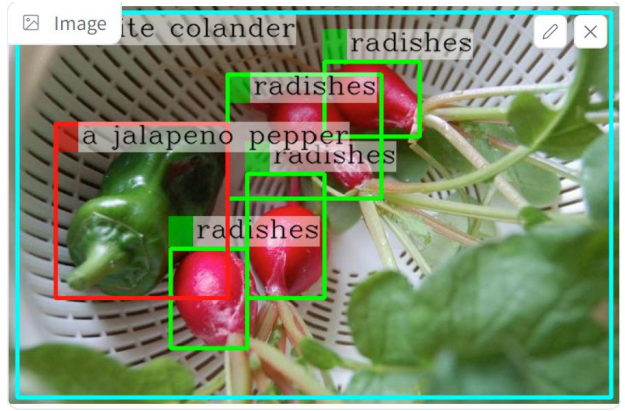
模型的训练致力于构建**开放的概念空间**，目标检测不再受限于带标注数据的少数类别，从而实现更加泛化的目标检测，识别出更多的未知物体类别



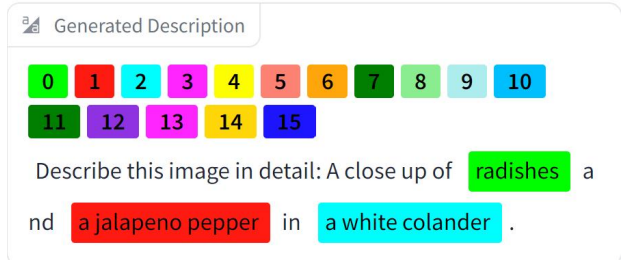
开放视觉感知

传统方法

开放视觉感知

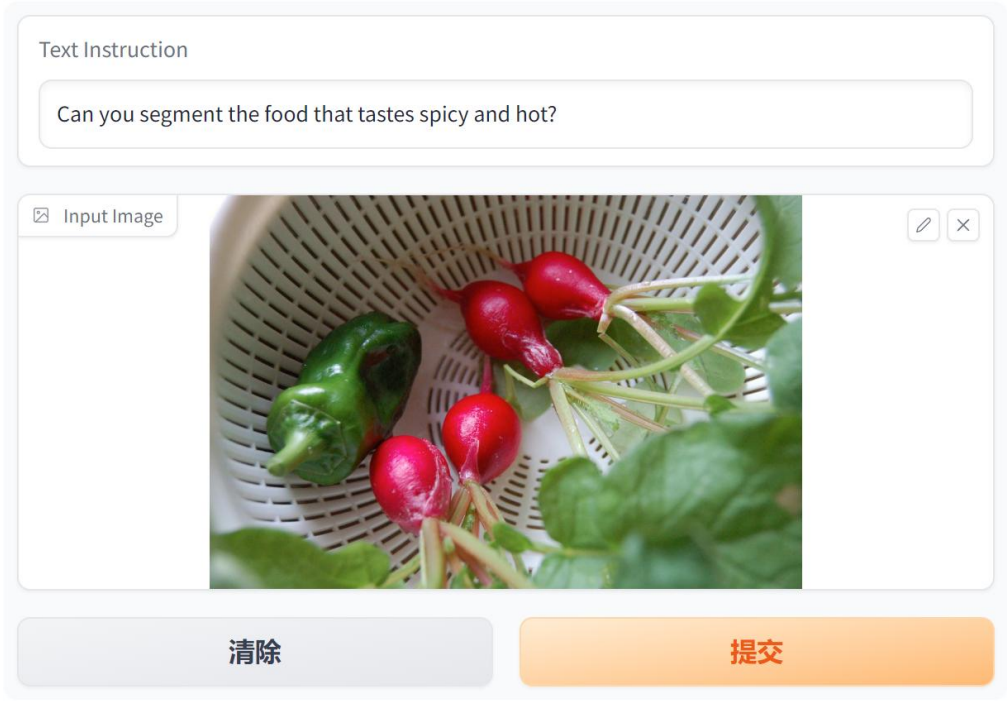


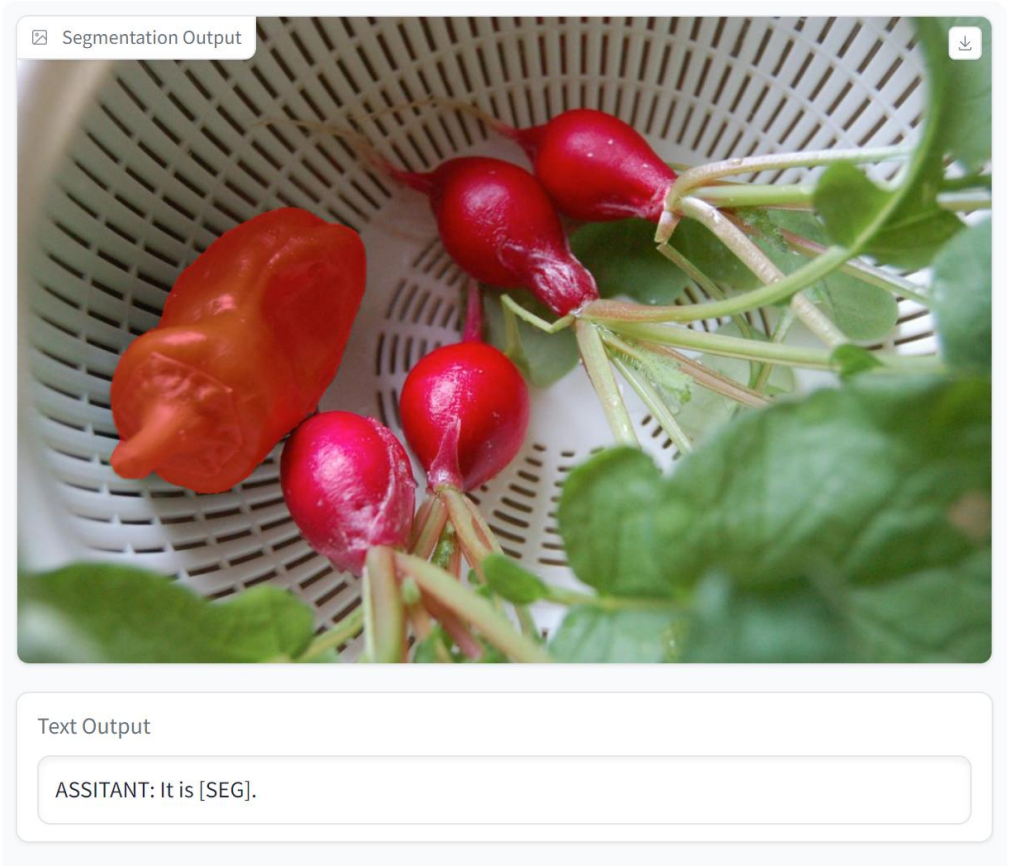
模型的训练致力于构建**开放的概念空间**，目标检测不再受限于带标注数据的少数类别，从而实现更加泛化的目标检测，识别出更多的未知物体类别



开放视觉感知

具备对语义概念的理解能力，因此可以基于模型的现实世界知识储备，处理复杂的自然语言指令，并给出更精细的分割结果





开放视觉感知

具备对语义概念的理解能力，因此可以基于模型的现实世界知识储备，处理复杂的自然语言指令，并给出更精细的分割结果

Text Instruction

What can indicate the country where this picture is taken. Please output segmentation mask and explain why.

Input Image

清除　　提交

Segmentation Output

Text Output

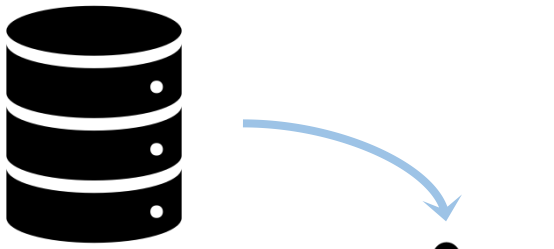
ASSITANT: Sure, [SEG]. In the image, the presence of a flag with a red and white pattern can indicate that the picture is taken in a country with a flag that has those colors. One possibility is that the flag represents Switzerland, as it is known for having a red and white flag. However, without more context or information about the location, it is not possible to confirm the exact country where the picture was taken.

02

研究现状

研究现状

从数据获取知识

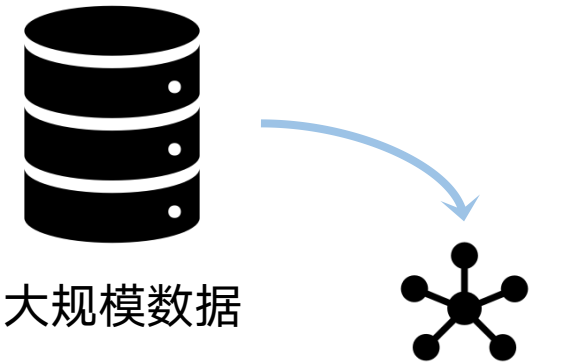


记忆事实知识

概念理解能力

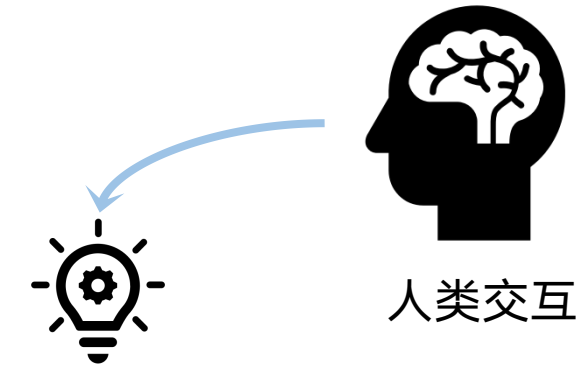
研究现状

从数据获取知识，以交互引入概念



记忆事实知识

概念理解能力

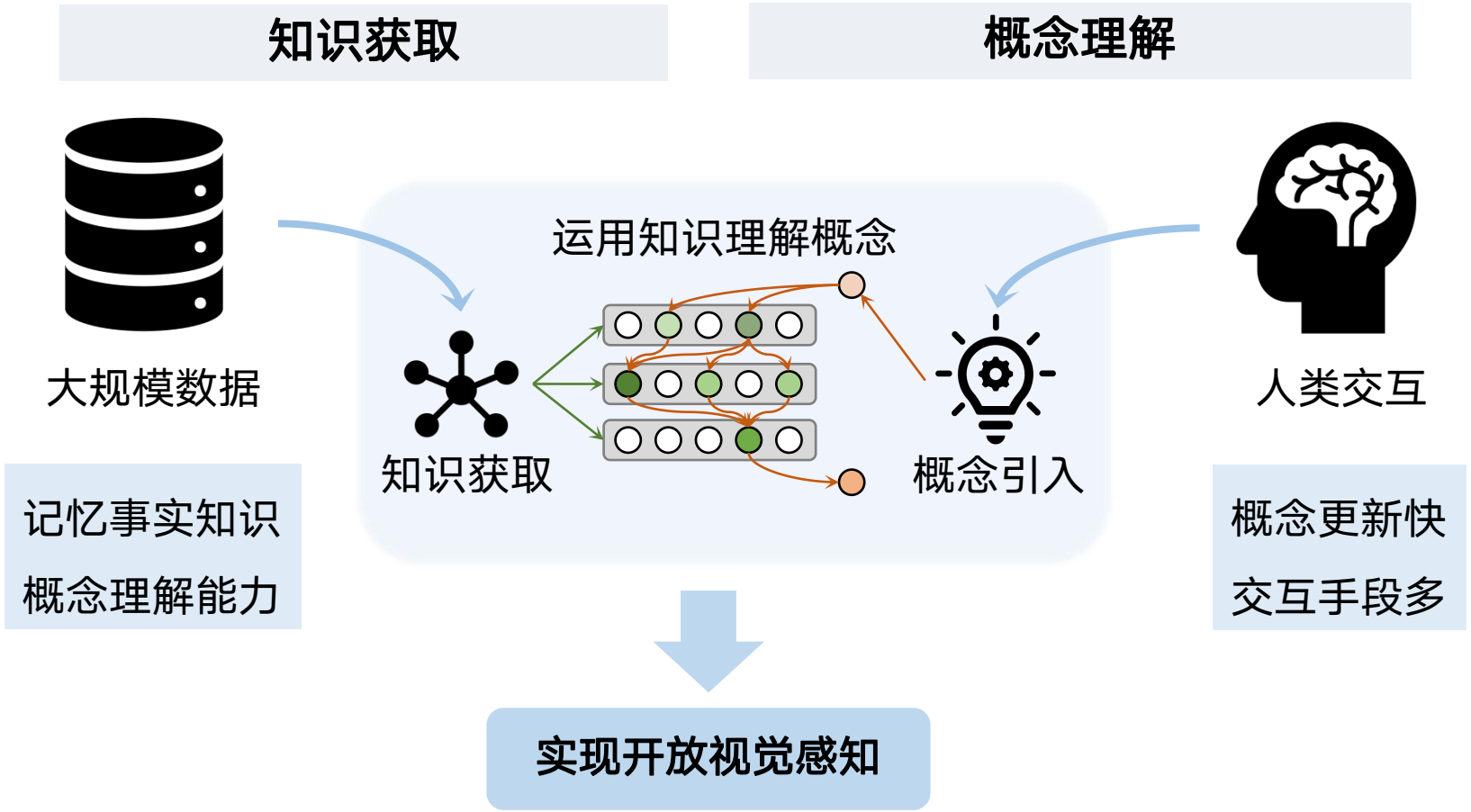


概念更新快

交互手段多

研究现状

从数据获取知识，以交互引入概念，实现开放感知

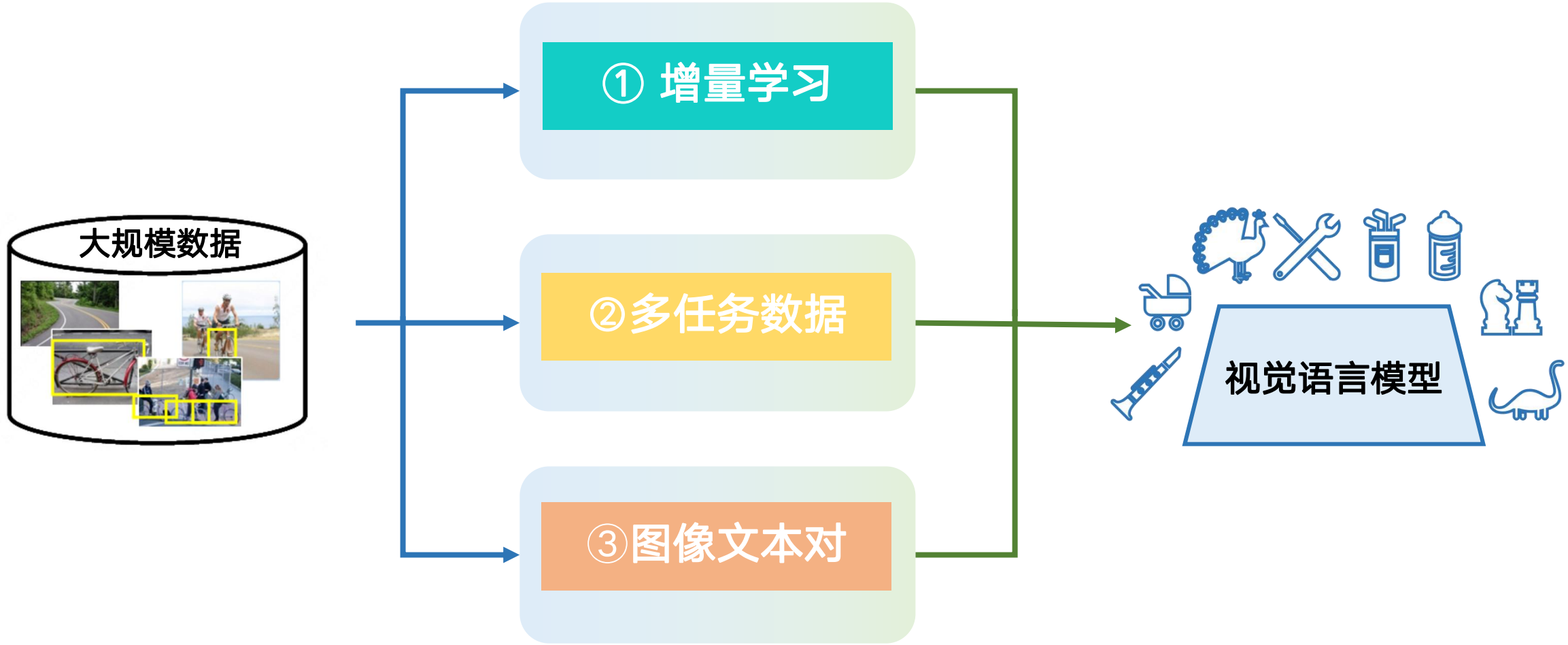


研究现状-知识获取

从大规模数据**提炼**知识，并**迁移**到下游感知任务

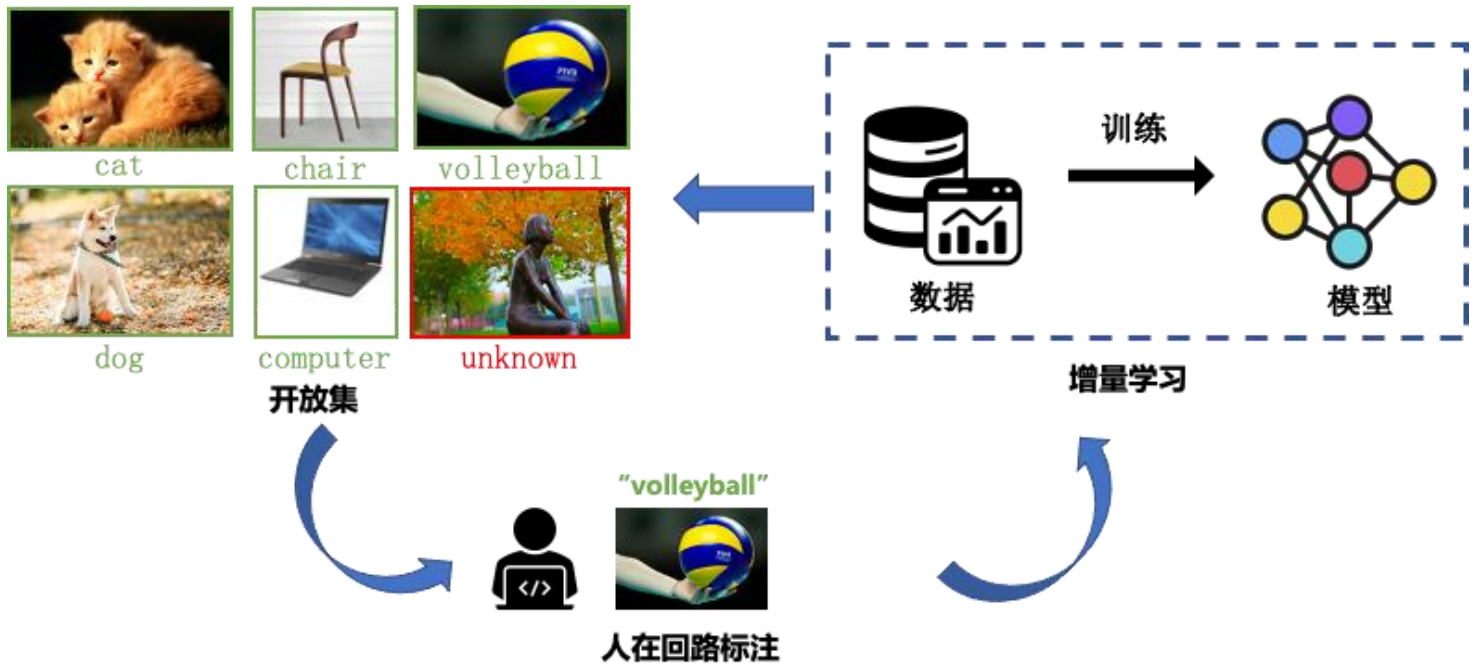
研究现状-知识获取

从多种数据源提炼知识



研究现状-知识获取-增量学习

主动发现未知概念，以人在回路的方式标注新概念，通过增量学习使模型理解新概念



特点：

- 理解新概念周期长
- 标注质量高
- 标注成本高

Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.
Jiyang Zheng, et al. “Towards open-set object detection and discovery.” CVPR 2022.

研究现状-知识获取-增量学习

Towards open world object detection认为模型应做到：

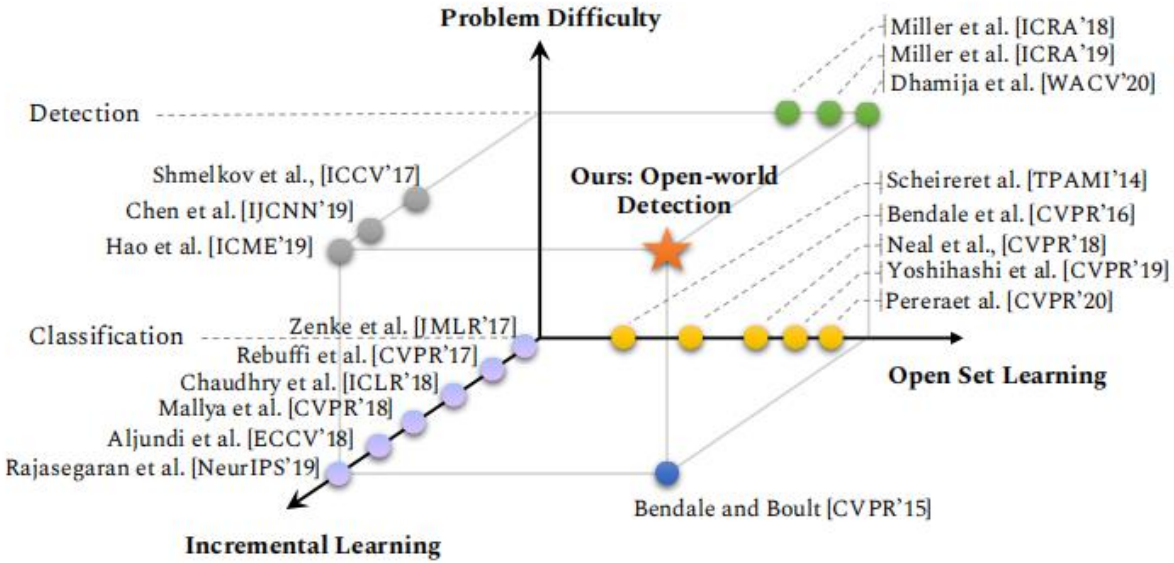
1. 在没有明确监督的情况下能够将没有见过的目标识别为“未知”类；
2. 对于认定为“未知”的类，当逐步获得对应的标签之后，能够增量地学习它们，且不遗忘以前学过的类别。

这类任务被定义为开放世界目标检测（Open World Object Detection，OWOD）任务；

并提出网络模型ORE来解决该任务。

特点：

- OWOD可以更紧密地模拟现实世界
- ORE 是第一个直面 OWOD 的方法

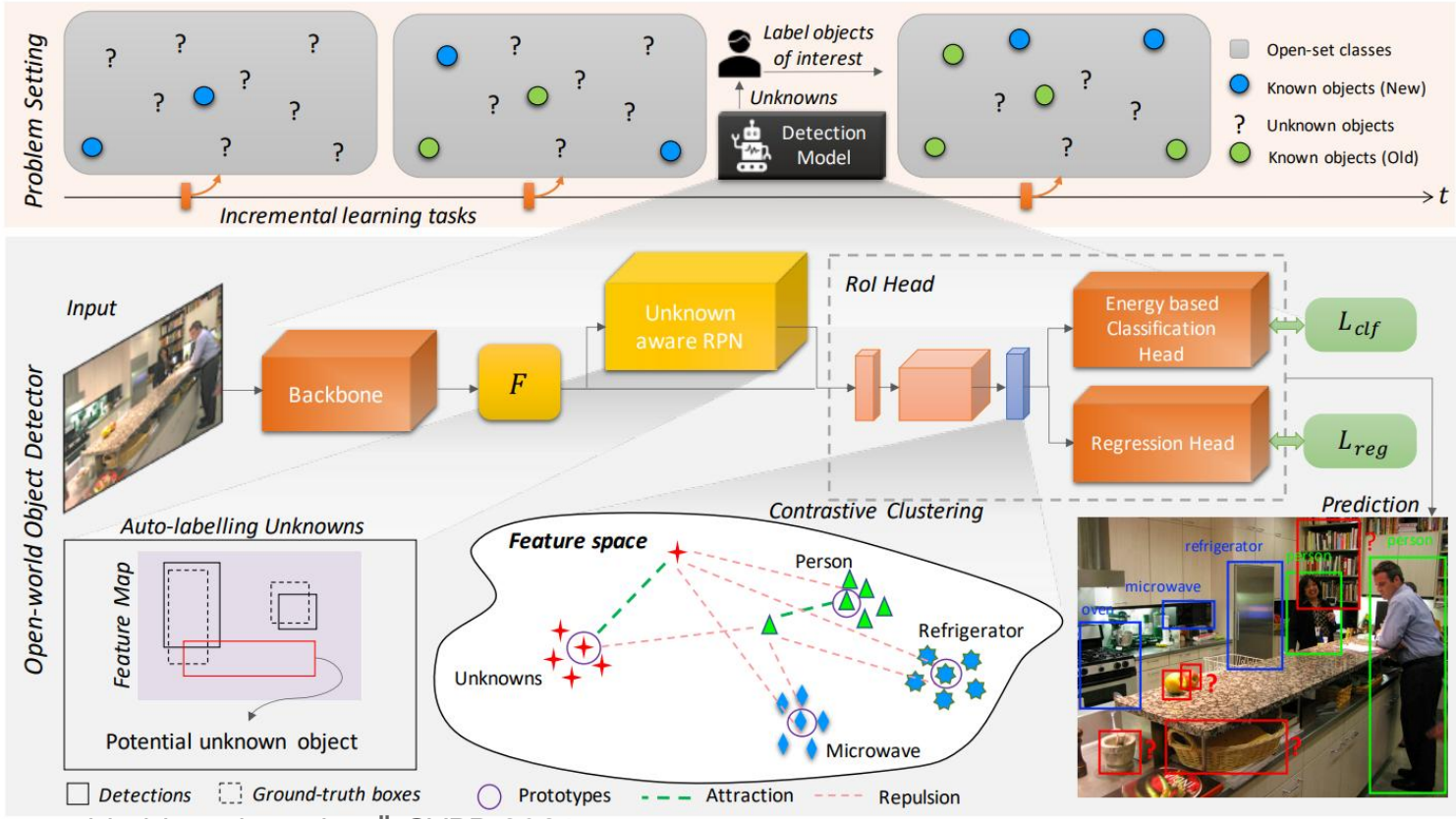


Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

研究现状-知识获取-增量学习-ORE

ORE是一个基于 two-stage Faster RCNN pipeline 的开放世界目标检测器。

- 主要思路：在隐空间中聚类，把不属于当前已知类别的特征分类为unknown
- 使用基于能量的分类头和unknown-aware的RPN来识别潜在的未知目标



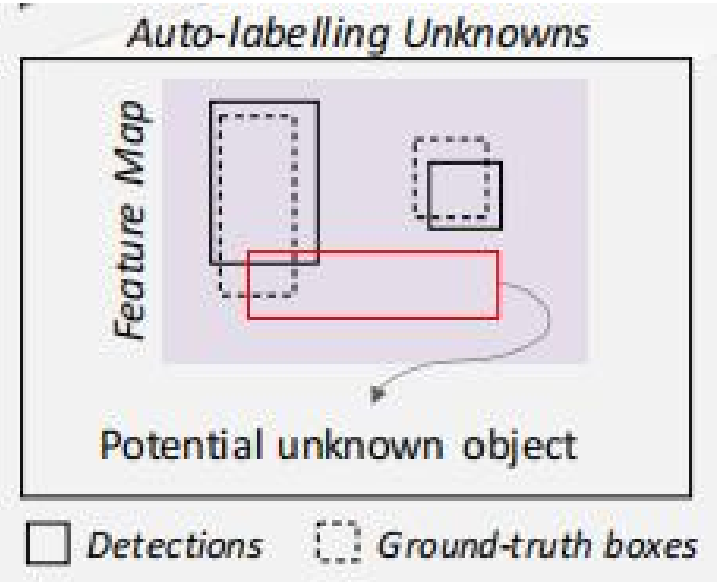
Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

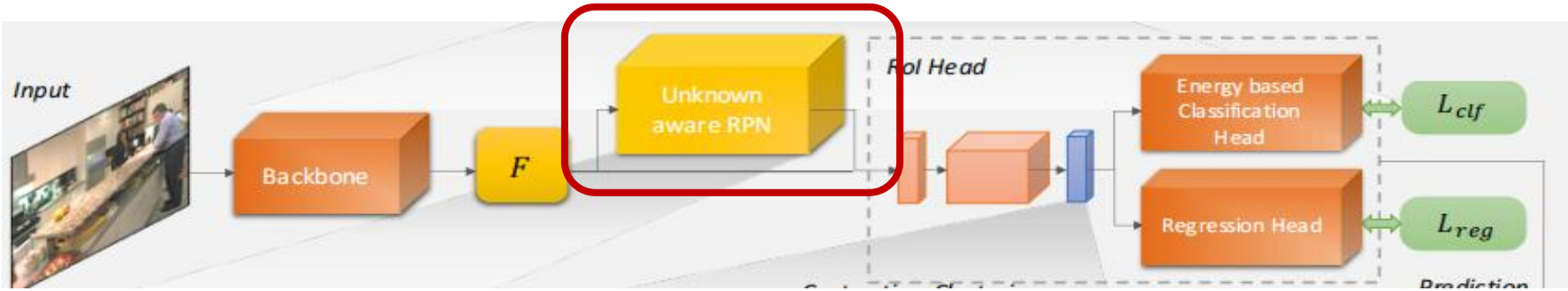
研究现状-知识获取-增量学习-ORE

RPN自动标注机制

- 我们无法提前标注出未知类别的目标
- RPN产生的候选框只区分前景和背景

所以，我们认为背景框中分数较高的就是没有标注的目标，直接将分数较高的背景框视为未知类别目标。将其作为未知目标的建议框向后传递，实现自动标注





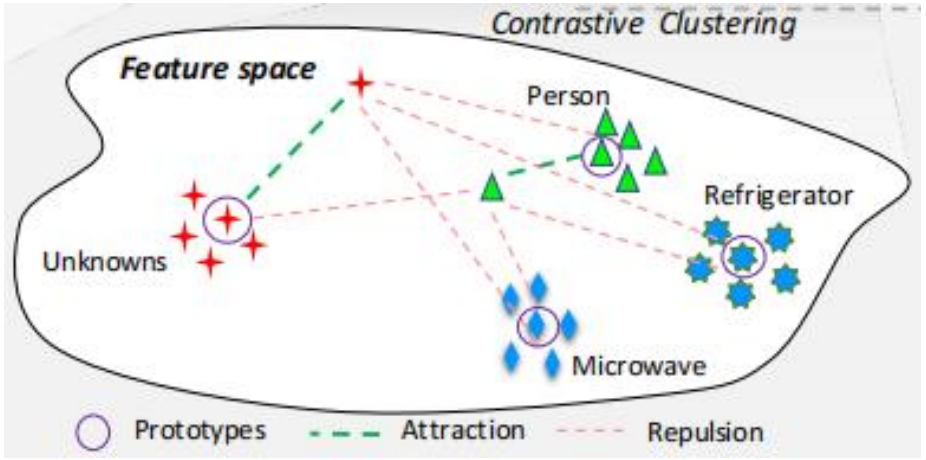
Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

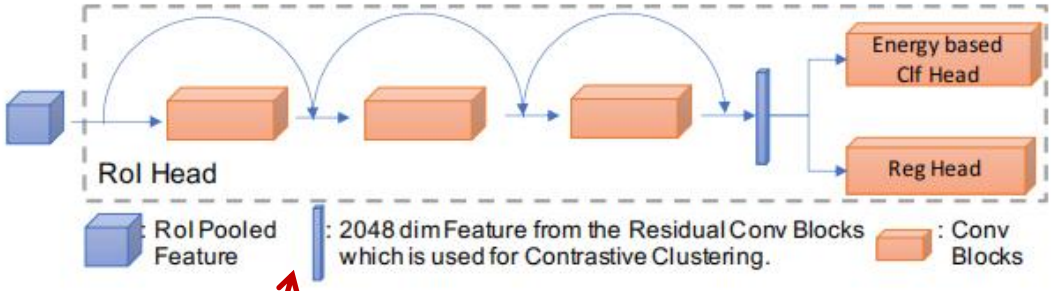
研究现状-知识获取-增量学习-ORE

对比聚类

为了实现“潜在空间中的类分离”的一种方法是将其建模为一个对比聚类问题：

- 使模型能够将未知实例与已知实例分开聚类，从而增强未知识别；
- 确保每个类的实例与其他类很好地分离，缓解了遗忘问题。





Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

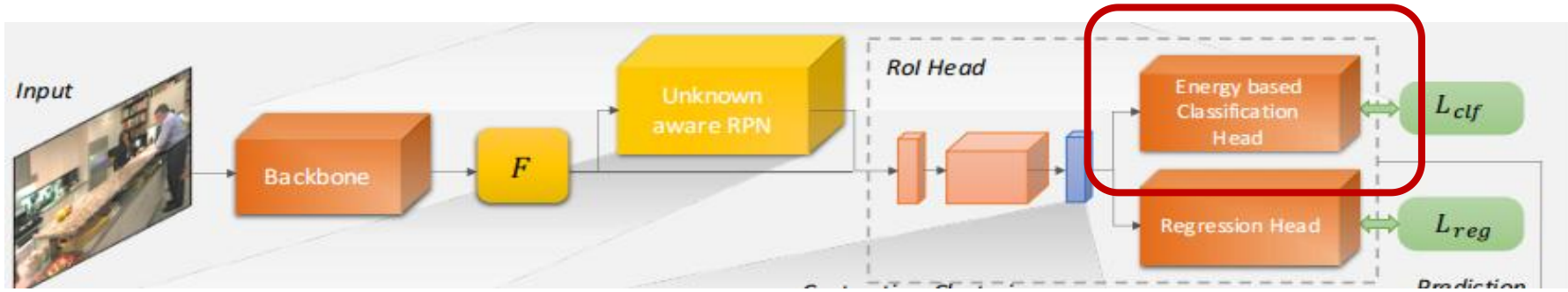
研究现状-知识获取-增量学习-ORE

基于能量的分类模型

在推测阶段，基于输入的特征向量*f*，利用一个基于能量的模型去计算，获取其预测的类别。

$$E(\boldsymbol{f}) = -T \log \int_{l'} \exp\left(-\frac{E(\boldsymbol{f}, l')}{T}\right)$$

通过能量值*E*判断是不是已知类别；若为已知类别，再看哪个类别的输出值更高



Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

研究现状-知识获取-增量学习-ORE

开放世界评估协议：数据分割

根据数据类别分成一系列的任务：

$$T = \{T_1; \ldots\ldots; T_t; \ldots\ldots\}$$

某个任务T_t（Task t）的所有类将在时间点t被引入系统。

学习T_t时，时间点t之前任务的所有类将被视为已知，T_t之后任务包含的类别将被视为未知。

	Task 1	Task 2	Task 3	Task 4
Semantic split	VOC Classes	Outdoor, Accessories, Appliance, Truck	Sports, Food	Electronic, Indoor, Kitchen, Furniture
# training images	16551	45520	39402	40260
# test images	4952	1914	1642	1738
# train instances	47223	113741	114452	138996
# test instances	14976	4966	4826	6039

Table 1: The table shows task composition in the proposed Open World evaluation protocol. The semantics of each task and the number of images and instances (objects) across splits are shown.

Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

研究现状-知识获取-增量学习-ORE

开放世界评估协议：评估指标

由于未知目标很容易与已知目标混淆，因此除mAP外，还定义一个评价指标：

Wilderness Impact （WI）。其中$P_{\mathcal{K}}$是指在已知类别上评估时模型的精度，$P_{\mathcal{K}\cup\mathcal{U}}$是在已知和未知类别上评估时的精度。

理想情况下，WI应该更小，因为当未知对象被添加到测试集时，精度不能下降。

$$\text{Wilderness Impact}\,(WI) = \frac{P_{\mathcal{K}}}{P_{\mathcal{K}\cup\mathcal{U}}} - 1$$

除了WI之外，还使用Absolute Open-Set Error（A-OSE）来报告被错误分类为任何已知类的未知对象的数量。

Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

研究现状-知识获取-增量学习-ORE

WI和A-OSE量化了模型如何处理未知类，而平均精度均值（mAP）则衡量其检测已知类的效果。ORE在所有指标上始终优于基于 Faster-RCNN based baseline。

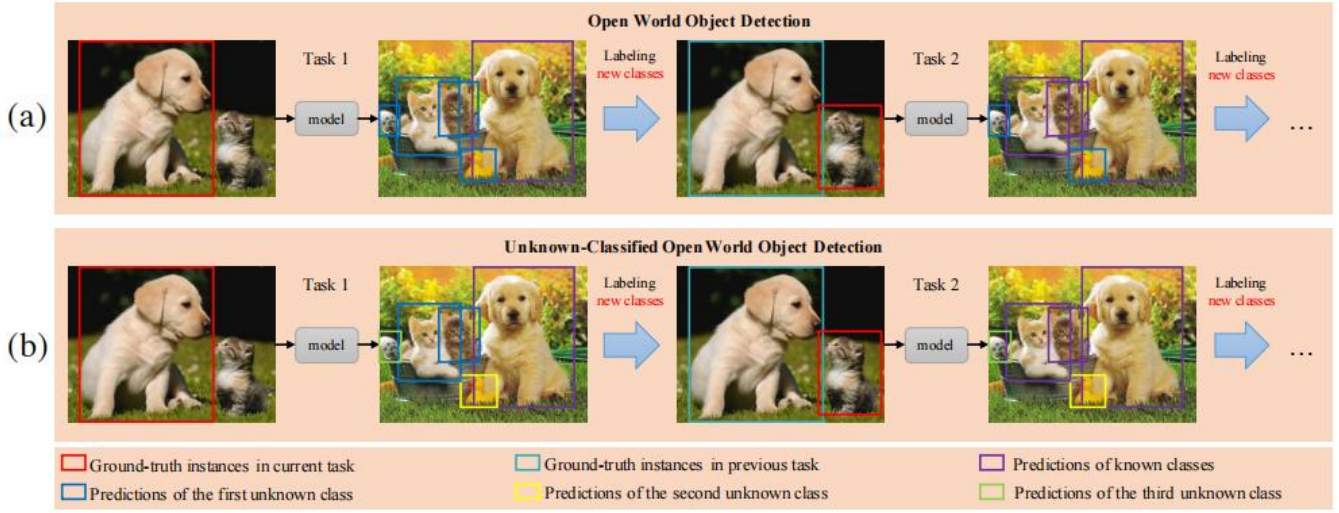
Task IDs (→)	Task 1			Task 2					Task 3					Task 4		
	WI	A-OSE	mAP (↑)	WI	A-OSE	mAP (↑)			WI	A-OSE	mAP (↑)			mAP (↑)		
	(↓)	(↓)	Current known	(↓)	(↓)	Previously known	Current known	Both	(↓)	(↓)	Previously known	Current known	Both	Previously known	Current known	Both
Oracle	0.02004	7080	57.76	0.0066	6717	54.99	30.31	42.65	0.0038	4237	40.23	21.51	30.87	32.52	19.27	31.71
Faster-RCNN	0.06991	13396	56.16	0.0371	12291	4.076	25.74	14.91	0.0213	9174	6.96	13.481	9.138	2.04	13.68	4.95
Faster-RCNN + Finetuning	Not applicable as incremental component is not present in Task 1			0.0375	12497	51.09	23.84	37.47	0.0279	9622	35.69	11.53	27.64	29.53	12.78	25.34
ORE	**0.02193**	**8234**	**56.34**	**0.0154**	**7772**	52.37	25.58	**38.98**	**0.0081**	**6634**	37.77	12.41	**29.32**	30.01	13.44	**26.66**

Joseph, K. J., et al. “Towards open world object detection.” CVPR 2021.

研究现状-知识获取-增量学习-UC OWOD

UC-OWOD

OWOD将未知对象检测为同一个类：“unknown”。UC-OWOD任务目标是检测到未知对象作为不同的类。



特点：

- 将未知实例区分为多个未知类
- 设计了新的评估协议

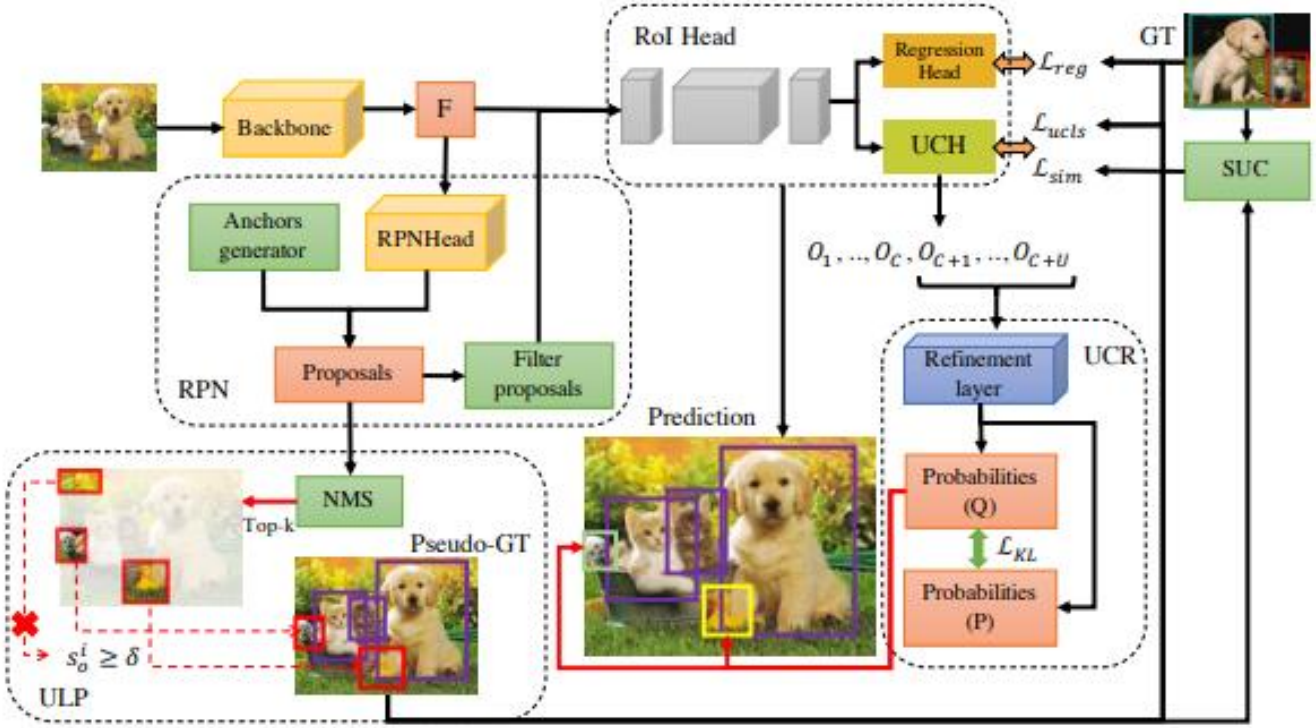
Zhiheng Wu, Yue Lu, et al. “UC-OWOD: Unknown-Classified Open World Object Detection.” ECCV 2022

研究现状-知识获取-增量学习-UC OWOD

使用faster-RCNN作为基础检测器。

引入：

- ULP和UCH来解决从背景中发现未知类的问题，
- SUC将未知对象检测为不同的类，
- UCR来细化未知目标的分类，增强算法的鲁棒性。



Zhiheng Wu, Yue Lu, et al. “UC-OWOD: Unknown-Classified Open World Object Detection.” ECCV 2022

研究现状-知识获取-增量学习-UC OWOD

评价指标

mAP不可以评估多个未知类别。UC-OWOD引入了一种新的评价度量：未知平均精度均值（UC-mAP）来评价未知类的检测。

$$\text{UC-mAP}(\mathcal{Y}_{gt}, \mathcal{Y}_{pre}) = \max_{perm \in \mathcal{P}} \text{mAP}(perm(\mathcal{Y}_{pre}), \mathcal{Y}_{gt}),$$

其中*perm* $\in P$是1到U中所有排列的集合，U是未知类的数量，y_{pre}是预测值，y_{gt}是基本真实值。

保留OWOD的评价指标WI与A-OSE，同时也使用最大匹配后的未知类召回率UC-Recall作为评估度量。

Zhiheng Wu, Yue Lu, et al. “UC-OWOD: Unknown-Classified Open World Object Detection.” ECCV 2022

研究现状-知识获取-增量学习-UC OWOD

同时也保留**OWOD**的评价指标**WI**与**A-OSE**，实验效果优于**OWOD**任务的**ORE**方法

Task 1		Oracle	Faster-RCNN	Faster-RCNN +Finetuning	ORE	Ours	Ours+UCR
UC-mAP	(↑)	0	0	-	0.0133	**0.1344**	-
WI	(↓)	-	0.0188	-	0.0155	**0.0136**	-
A-OSE	(↓)	-	13300	-	10672	**9294**	-
UC-Recall	(↑)	-	0	-	0.7772	**2.3915**	-

Task 2		Oracle	Faster-RCNN	Faster-RCNN +Finetuning	ORE	Ours	Ours+UCR
UC-mAP	(↑)	15.50	0	0	0.0065	0.0862	**0.1694**
WI	(↓)	0.0022	0.0069	0.0140	0.0153	0.0116	0.0117
A-OSE	(↓)	6050	4582	7169	10376	5602	5602
UC-Recall	(↑)	40.45	0	0	0.0371	2.6926	**3.4431**

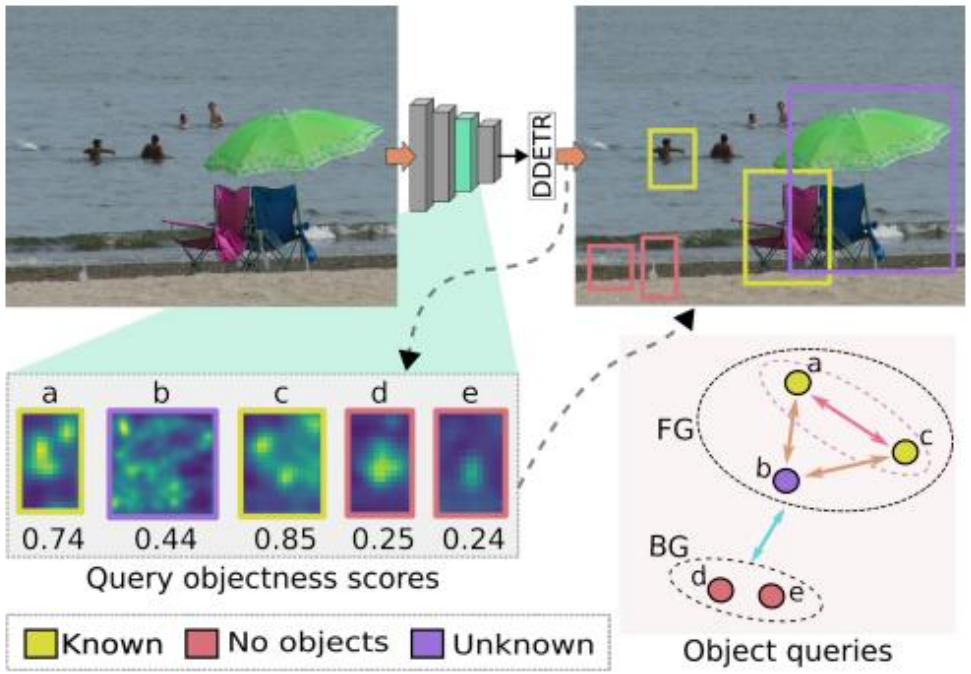
Task 3		Oracle	Faster-RCNN	Faster-RCNN +Finetuning	ORE	Ours	Ours+UCR
UC-mAP	(↑)	10.61	0	0	0.0070	0.0249	**0.0744**
WI	(↓)	0.0042	0.0241	0.0099	0.0086	0.0073	**0.0073**
A-OSE	(↓)	4857	4841	9181	7544	3801	**3801**
UC-Recall	(↑)	28.54	0	0	0.8833	4.8077	**8.7303**

Zhiheng Wu, Yue Lu, et al. “UC-OWOD: Unknown-Classified Open World Object Detection.” ECCV 2022

研究现状-知识获取-增量学习-OW DETR

OW-DETR

基于transformer的single-stage框架，假设对未知的情况不进行监督，这样更加接近于真正的开放世界场景。使用更宽的上下文建模和更加少的假设来解决开放世界目标检测问题。



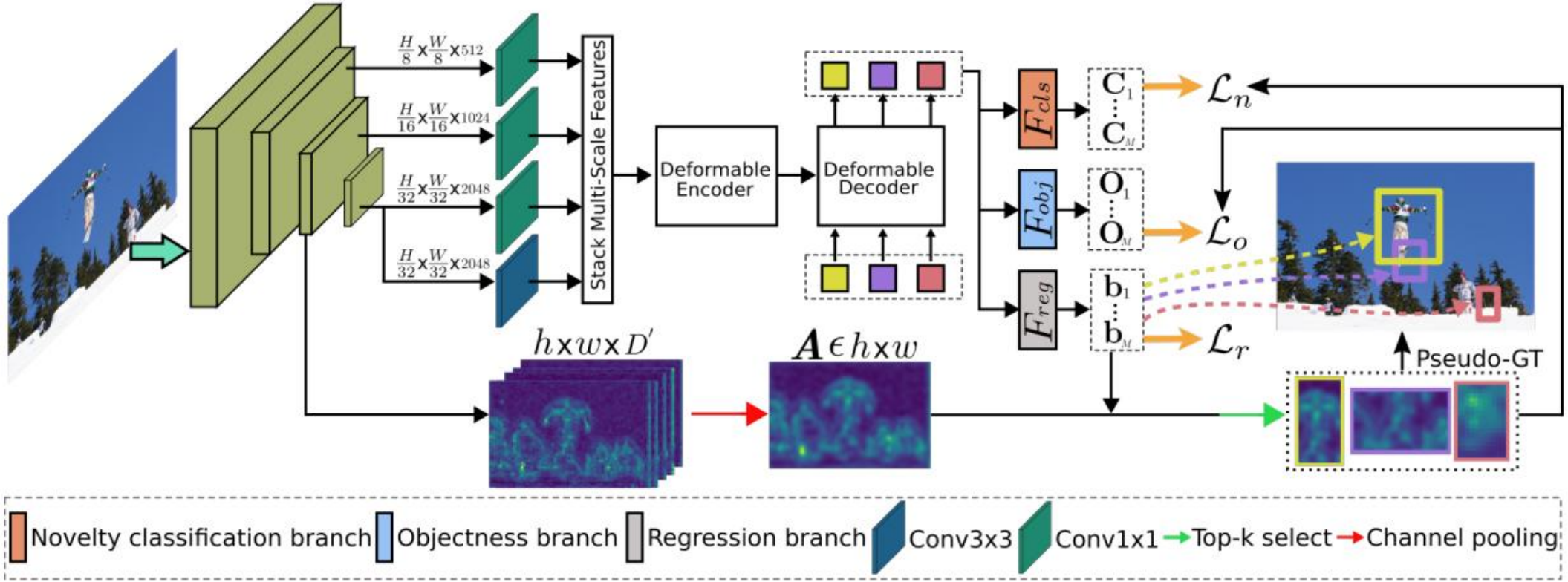
特点：

- 端到端
- 明确建模长程依赖关系
- 能更好地区分未知目标和背景

A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

基于视觉 **transformer**的多尺度上下文感知检测框架，具有专门的组件（注意力驱动的伪标签生成、新颖性分类、客观性评分）来处理开放世界任务，以有效地检测图像中的未知目标。

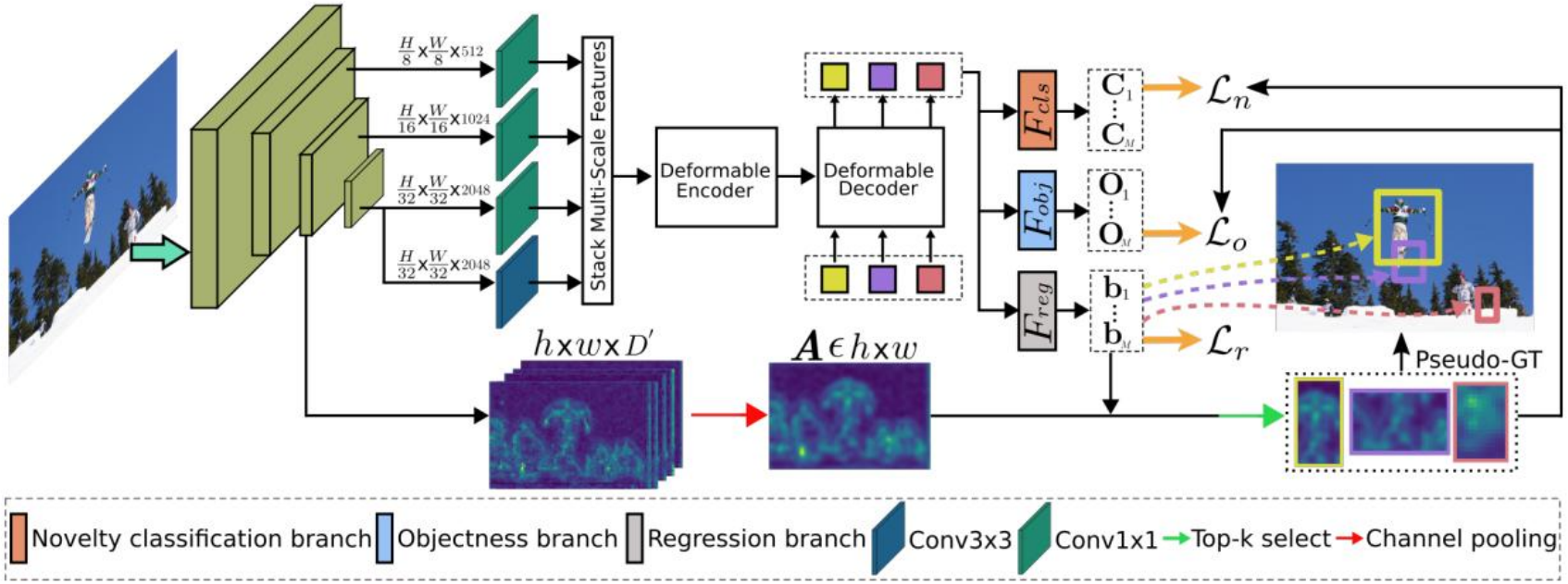


A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

首先通过一个多尺度backbone提取特征，随后输入到Transformer的Encoder中；

Decoder在跨尺度注意力和自注意机制驱动下，将一组可学习的query向量转换为对应的嵌入向量，随后被输入到三个独立分支（回归分支、新类别检测分支和客观性评分分支）进行后续的定位和识别。



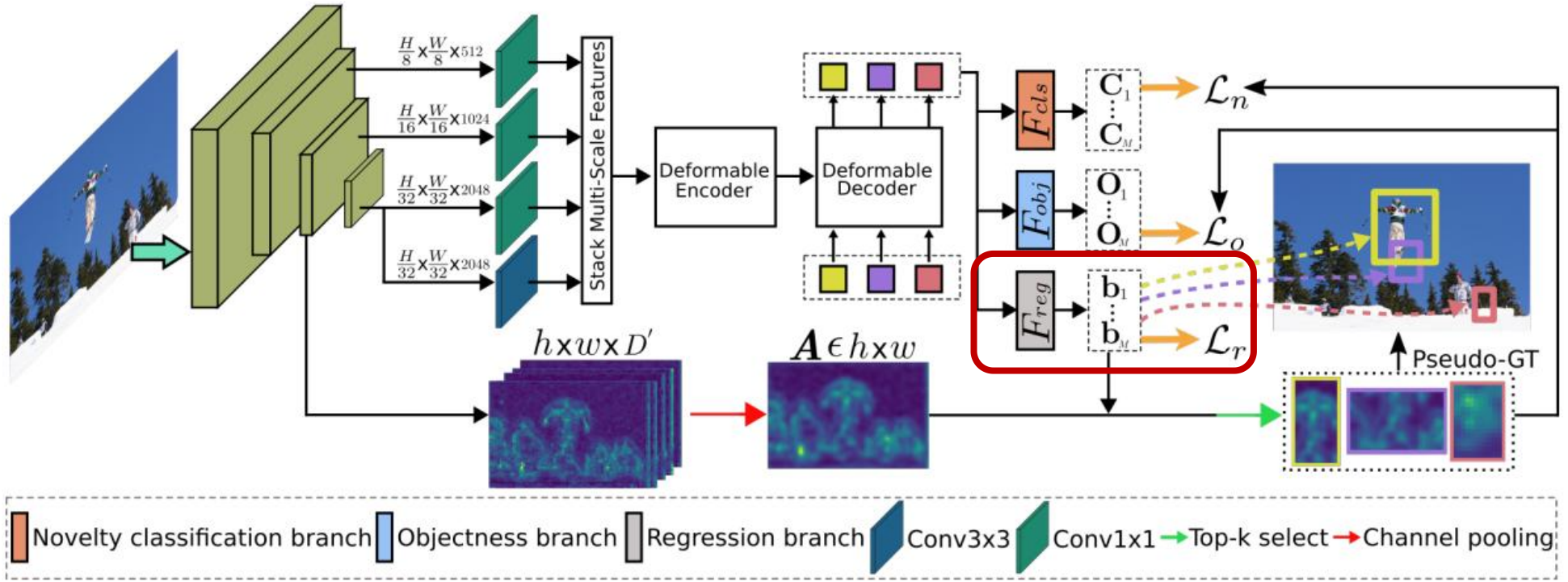
除这三个分支之外，本文方法的基本框架与Deformable DETR基本保持一致，首先使用二分匹配损失从GT标签中选择已知类的预测目标，然后从余下的目标查询向量中选择候选未知类的目标实例，其中候选目标实例是通过特征图的区域激活幅度来确定的， 较高对应的查询向量被标注为未知类别的伪标签数据。

A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

回归分支

首先使用二分匹配损失从GT标签中选择已知类的预测目标，然后通过注意力驱动的伪标签机制从余下的目标query向量中选择候选未知类的目标实例。

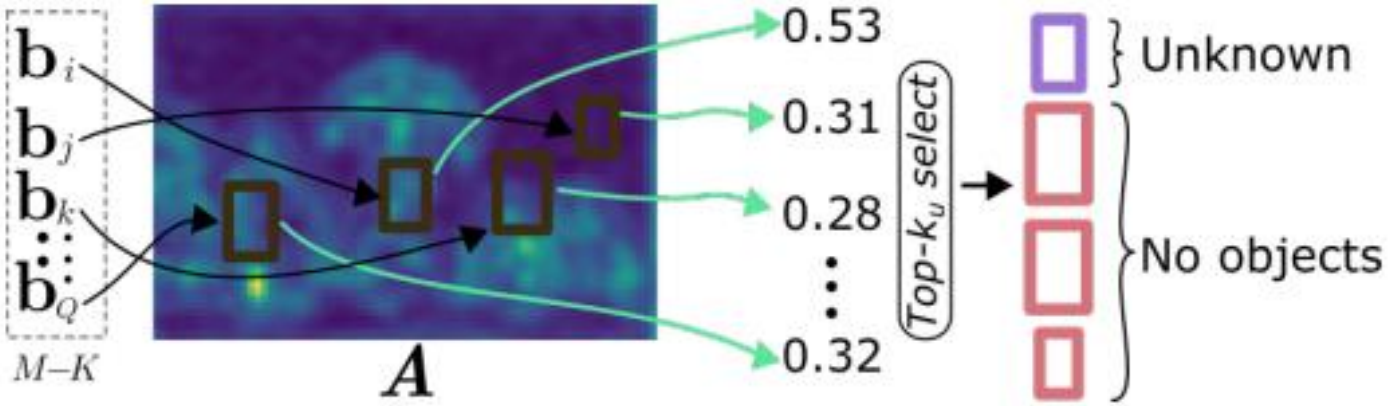


A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

注意力驱动的伪标签机制

输入图像经过**backbone**得到特征图，特征图上的各区域的激活值大小反映了该空间位置上存在目标置信度，如下图所示。从中选取置信度分数较高的实例标记为伪标签。



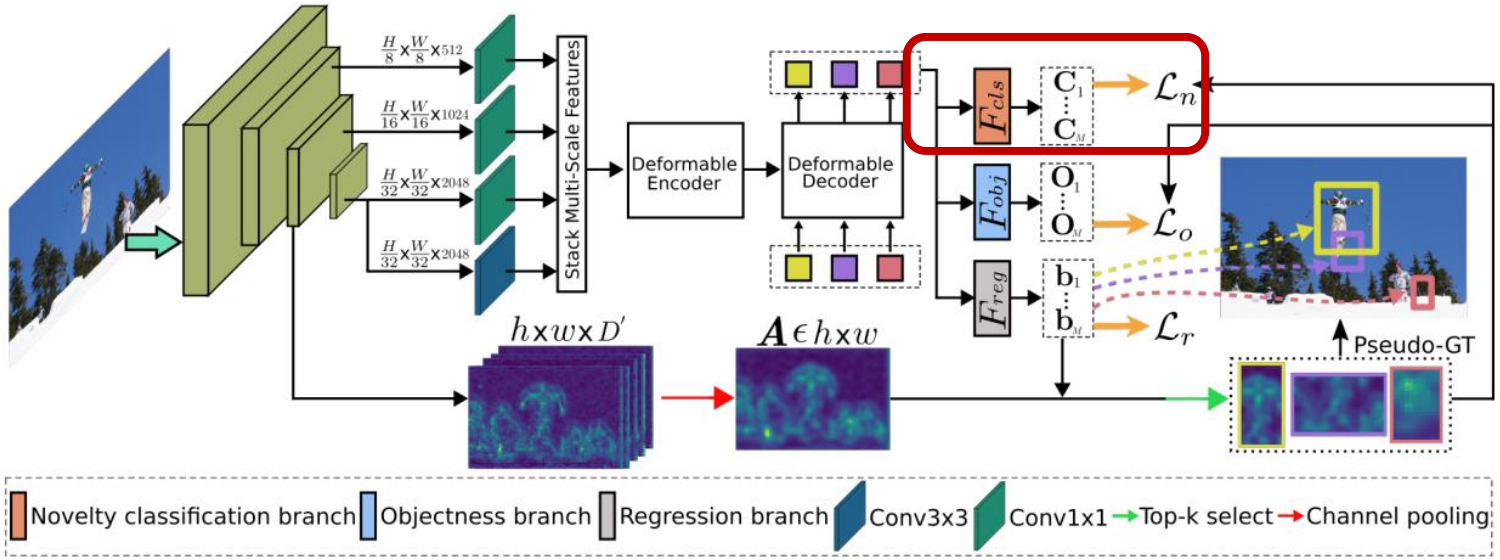
A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

新类别分类分支

标准目标检测器中的分类分支会将输入的query向量分类为已知类和背景类。但是，当遇到未知类别的目标时，这种检测器无法将其归入任何一种类别。

为了克服这一问题，在分类分支中引入了新类别标签（设置为0），训练数据为上一节得到的伪标签数据，其与已知类的实例共同训练分类分支。

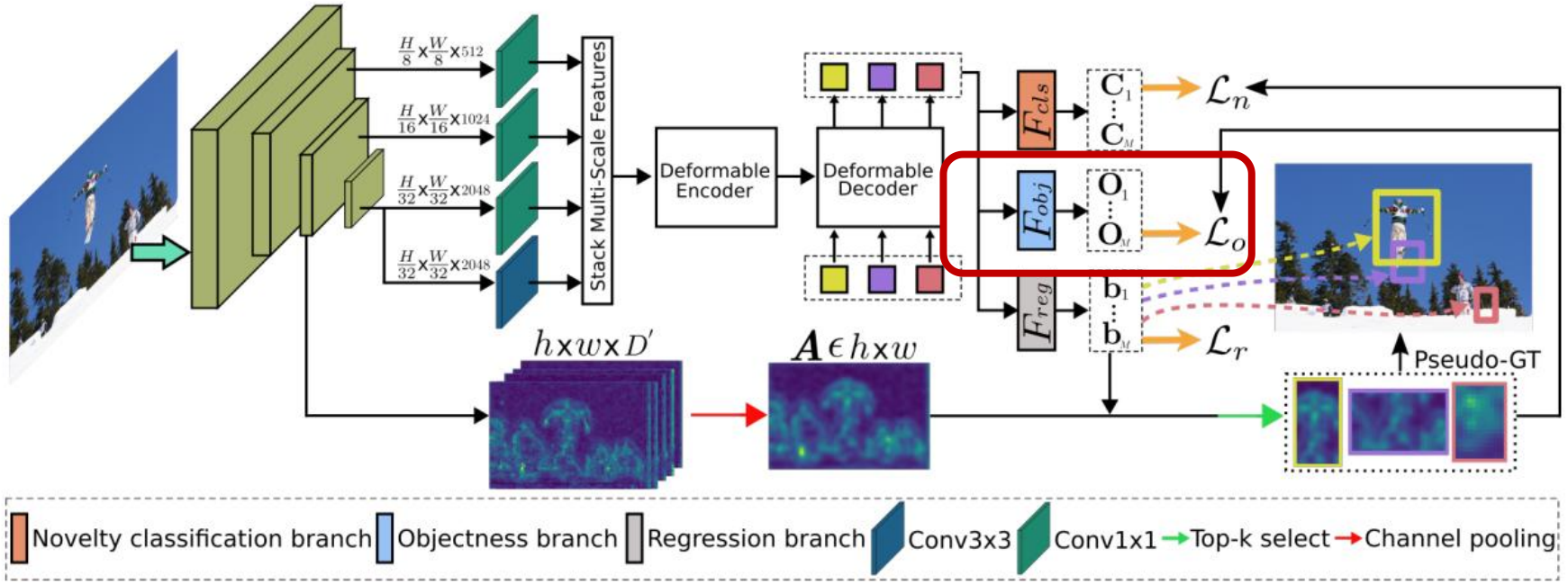


A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

客观性评分分支

引入一个前景目标分支 。该分支会对每个**query**向量给出一个客观性评分，以便更好的将前景目标（已知和未知）与背景分开。这种与类别无关的评分还有助于模型将知识从已知类别转移到未知类别。



A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-增量学习-OW DETR

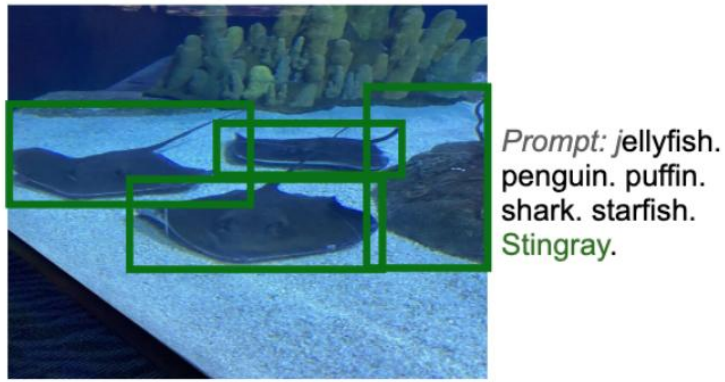
对于已知类别使用**mAP**作为评价指标。而对于未知类别使用召回率作为评价指标。与**ORE**方法比较了未知类别的检测效果，**OW-DETR**方法在跨任务上提高了**U-Recall**分数，展示了更强的未知类检测能力

Task IDs (→)	Task 1		Task 2				Task 3				Task 4			
	U-Recall	mAP (↑)	U-Recall	mAP (↑)			U-Recall	mAP (↑)			mAP (↑)			
	(↑)	Current known	(↑)	Previously known	Current known	Both	(↑)	Previously known	Current known	Both	Previously known	Current known	Both	
Faster-RCNN [32]	-	56.4	-	3.7	26.7	15.2	-	2.5	15.2	6.7	0.8	14.5	4.2	
Faster-RCNN + Finetuning	Not applicable in Task 1		-	51.0	25.0	38.0	-	38.2	13.6	30.0	29.7	13.0	25.6	
DDETR [38]	-	60.3	-	4.5	31.3	17.9	-	3.3	22.5	8.5	2.5	16.4	6.0	
DDETR + Finetuning	Not applicable in Task 1		-	54.5	34.4	44.8	-	40.0	17.8	33.3	32.5	20.0	29.4	
ORE – EBUI [15]	4.9	56.0	2.9	52.7	26.0	39.4	3.9	38.2	12.7	29.7	29.6	12.4	25.3	
Ours: OW-DETR	**7.5**	**59.2**	**6.2**	**53.6**	**33.5**	**42.9**	**5.7**	**38.3**	**15.8**	**30.8**	**31.4**	**17.1**	**27.8**	

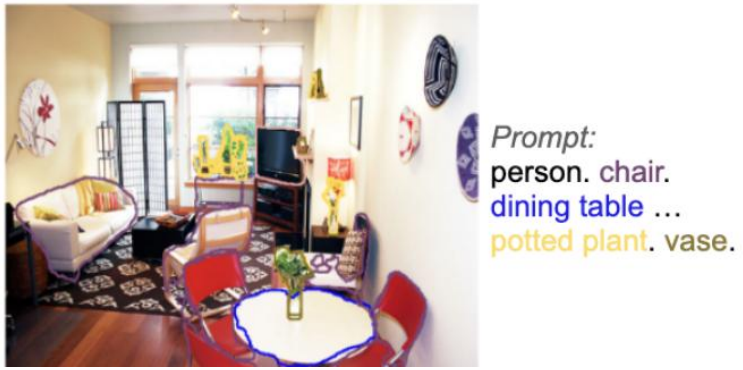
A.G., S.N., et al. “OW-DETR: Open-world Detection Transformer.” CVPR 2022

研究现状-知识获取-多任务数据

统一多任务数据的表示形式，联合训练模型以提炼知识



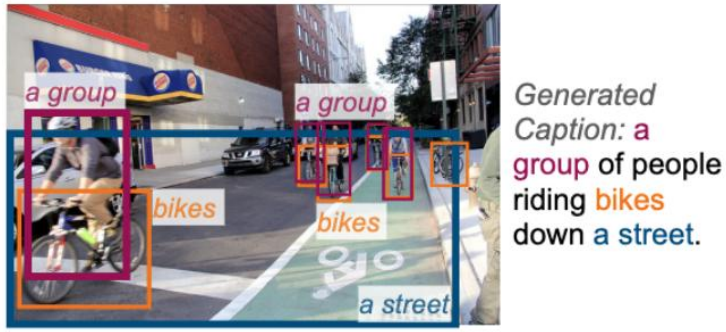
检测数据->grounding数据



分割数据->grounding数据

特点：

- 充分利用现有数据
- 标注质量高
- 感知粒度细



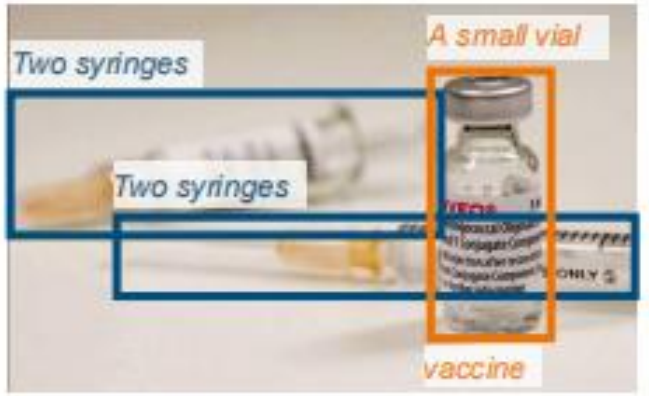
caption数据->grounding数据

VQA数据->grounding数据

研究现状-知识获取-多任务数据-GLIP

Grounded Language-Image Pre-training

GLIP模型，用于学习对象级、语言感知和语义丰富的视觉表征。GLIP将 object detection 和 phrase grounding 结合起来进行预训练。



Two syringes and a small vial of vaccine.

playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

优点：

- 通过将目标检测任务转化为grounding 任务
- 使用大量的图像-文本数据扩大视觉概念
- “大一统”的迁移学习模型

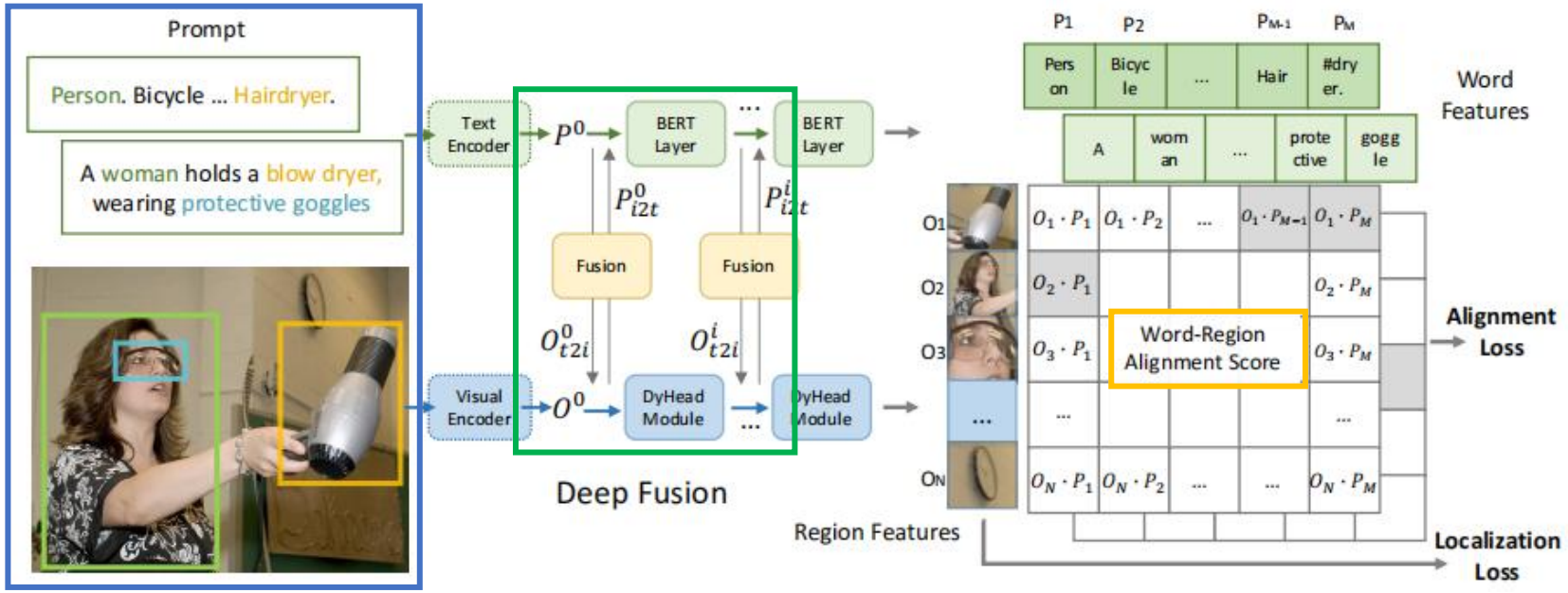
Li L H, et al. “Grounded language-image pre-training.” CVPR 2022.

研究现状-知识获取-多任务数据-GLIP

Grounded Language-Image Pre-training

将Detection和Grounding统一起来：

- 不仅将图像作为输入，还将描述检测任务中所有候选类别的文本提示作为输入
- 目标检测模型都可以通过用单词-区域对齐分数替换边框分类器中的目标分类logits而转换为Grounding模型
- 采用深度跨模态融合，这对于学习高质量的语言感知视觉表示和实现优秀的迁移学习性能至关重要



Li L H, et al. “Grounded language-image pre-training.” CVPR 2022.

研究现状-知识获取-多任务数据-GLIP

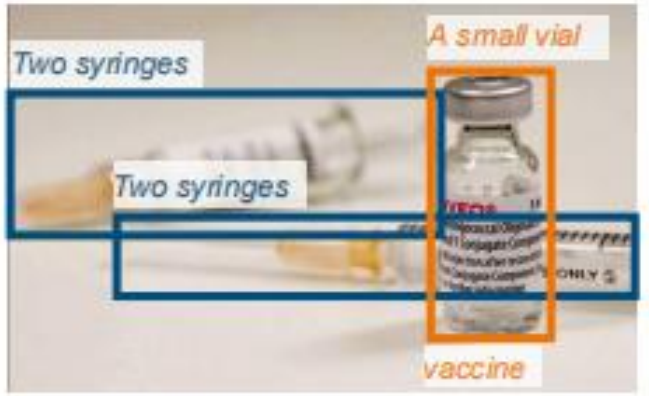
Grounded Language-Image Pre-training

大量的图像-文本数据扩大视觉概念：

- 使用Grounding模型（教师模型），通过self-training生成大量图像-文本配对数据的关联框来增加GLIP的预训练数据

- 教师模型可以准确地定位一些可能很难理解的概念，如注射器、疫苗、美丽的蓝绿色加勒比海，甚至是抽象的单词

- 语义丰富的数据上进行训练可以提供一个语义丰富的学生模型



Two syringes and a small vial of vaccine.

playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

Li L H, et al. “Grounded language-image pre-training.” CVPR 2022.

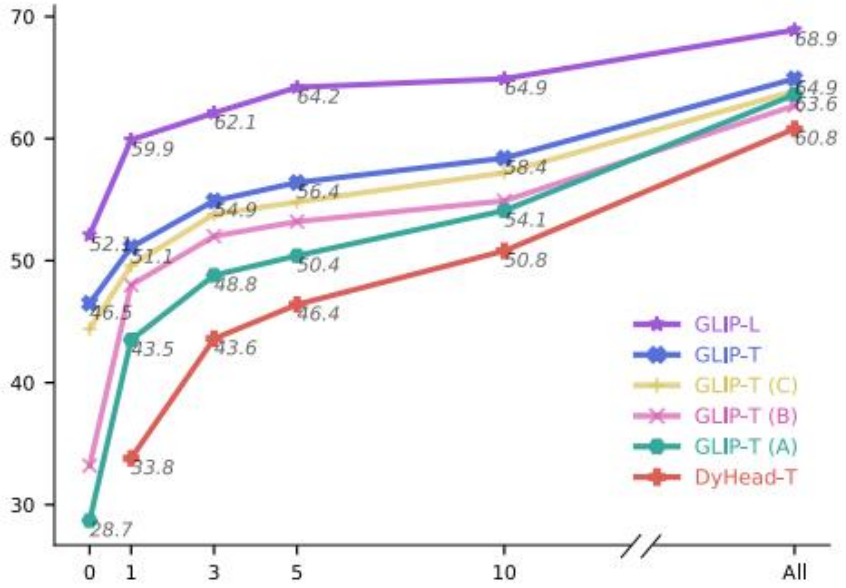
研究现状-知识获取-多任务数据-GLIP

Grounded Language-Image Pre-training

“大一统”的迁移学习模型：

- 关联重构和语义丰富的预训练有助于领域迁移
- GLIP可以迁移到各种任务上，几乎不需要额外的人工注释
- 当特定于任务的注释可用时，可以只调优特定于任务的提示嵌入，而不调整整个模型，同时保持模型参数不变，降低微调和部署的成本

Row	Pre-Training Data	COCO 2017val	LVIS MiniVal AP_r	AP_c	AP_f	AP
1	VG w/o COCO	26.9	4.9	10.4	23.2	16.1
2	+ GoldG	29.2	7.8	14.0	24.5	18.5
3	OpenImages	29.9	12.8	12.1	17.8	14.9
4	+ GoldG	33.6	15.2	16.9	24.5	20.4
5	O365	44.9	13.5	12.8	22.2	17.8
6	+GoldG	**46.7**	17.7	19.5	31.0	24.9
7	O365,GoldG,Cap4M	46.3	**20.8**	21.4	31.0	26.0
8	FourODs	46.3	15.0	**22.5**	**32.8**	**26.8**



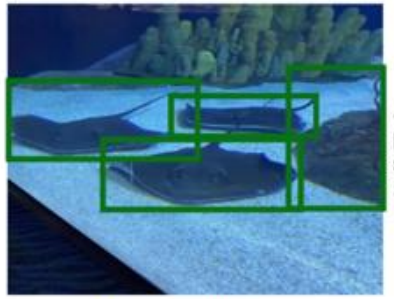
Li L H, et al. “Grounded language-image pre-training.” CVPR 2022.

研究现状-知识获取-多任务数据-GLIPv2

GLIPv2: Unifying Localization and VL Understanding

GLIPv2既服务于定位任务(例如，目标检测，实例分割)，也服务于视觉语言(VL)理解任务(例如VQA，image captioning)，通过三个预训练任务将定位预训练和视觉语言预训练(VLP)巧妙地统一起来

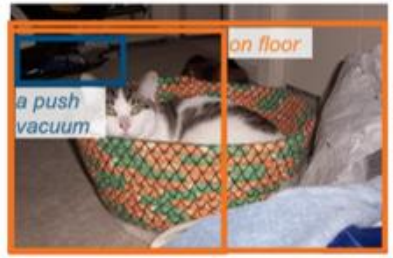
- phrase grounding作为检测任务的VL重构（GLIP）
- region-word contrastive learning作为一种新的region-word level对比学习任务
- masked language modeling(MLM)



检测数据->grounding数据

分割数据->grounding数据

caption数据->grounding数据



VQA数据->grounding数据

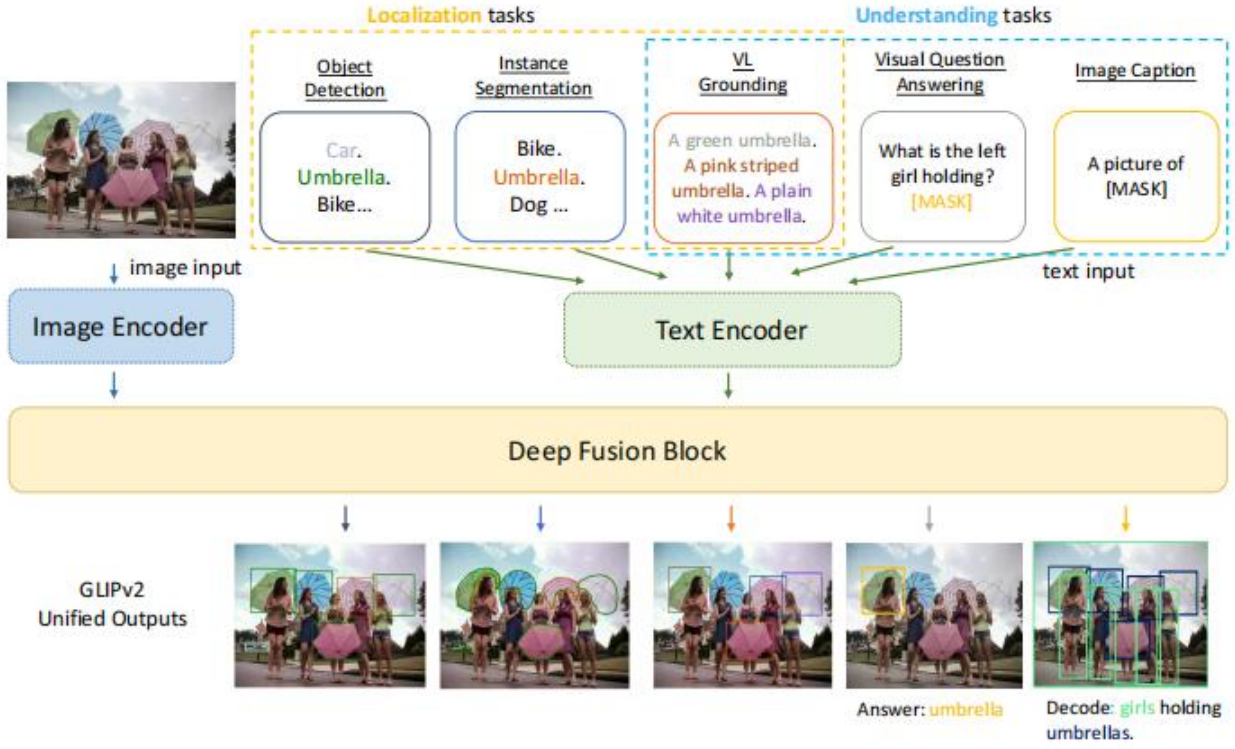
Haotian Zhang, et al. “GLIPv2: Unifying Localization and Vision-Language Understanding.” NeurIPS 2022.

研究现状-知识获取-多任务数据-GLIPv2

GLIPV1提出了一种detection转grounding的reformulation，提供了一种vision-only模型从VL数据丰富的语义信息中获益的可能。V2直接将grounding定义为localization和understanding的基础能力。

将所有任务数据转化为grounded数据，对GLIPv2进行预训练，进行grounded VL understanding

- localization任务涉及localization和semantic classification，其中分类可以使用classification-to-matching技巧转换为VL理解问题
- localization数据相应转化为VL grounding数据
- 海量的VL理解数据(图像-文本对)可以很容易地通过self-training的方式转化为VL grounded数据



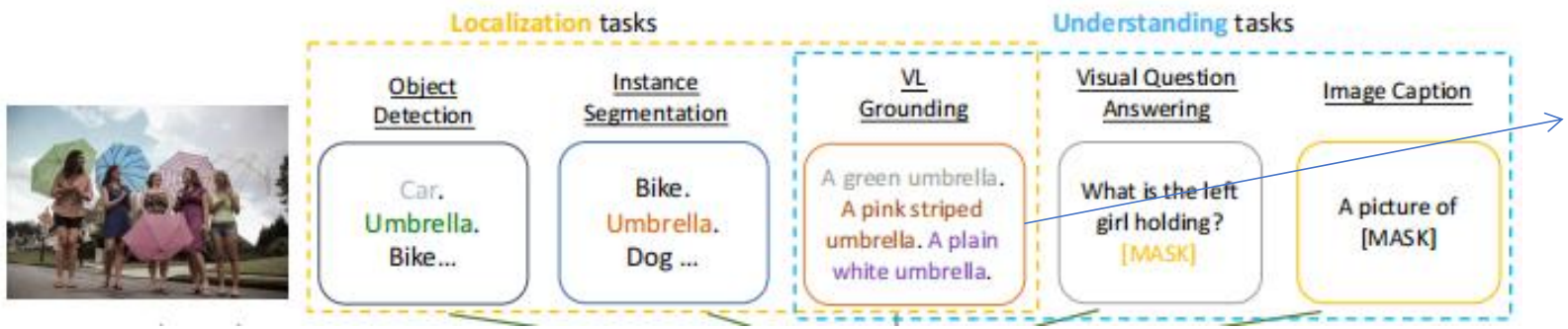
Haotian Zhang, et al. “GLIPv2: Unifying Localization and Vision-Language Understanding.” NeurIPS 2022.

研究现状-知识获取-多任务数据-GLIPv2

GLIPv2: Unifying Localization and VL Understanding

inter-image region-word contrastive learning：

- GLIP提出phrase grounding任务作为其预训练任务，这是一个简单的任务，没有充分利用数据信息
- 引入了新的inter-image region-word对比学习任务，该任务利用同一batch其他句子中的短语作为潜在负例，实现另一个更强大的VL grounding任务



丢失大量信息，其他颜色，其他类别等，可作为潜在负样本

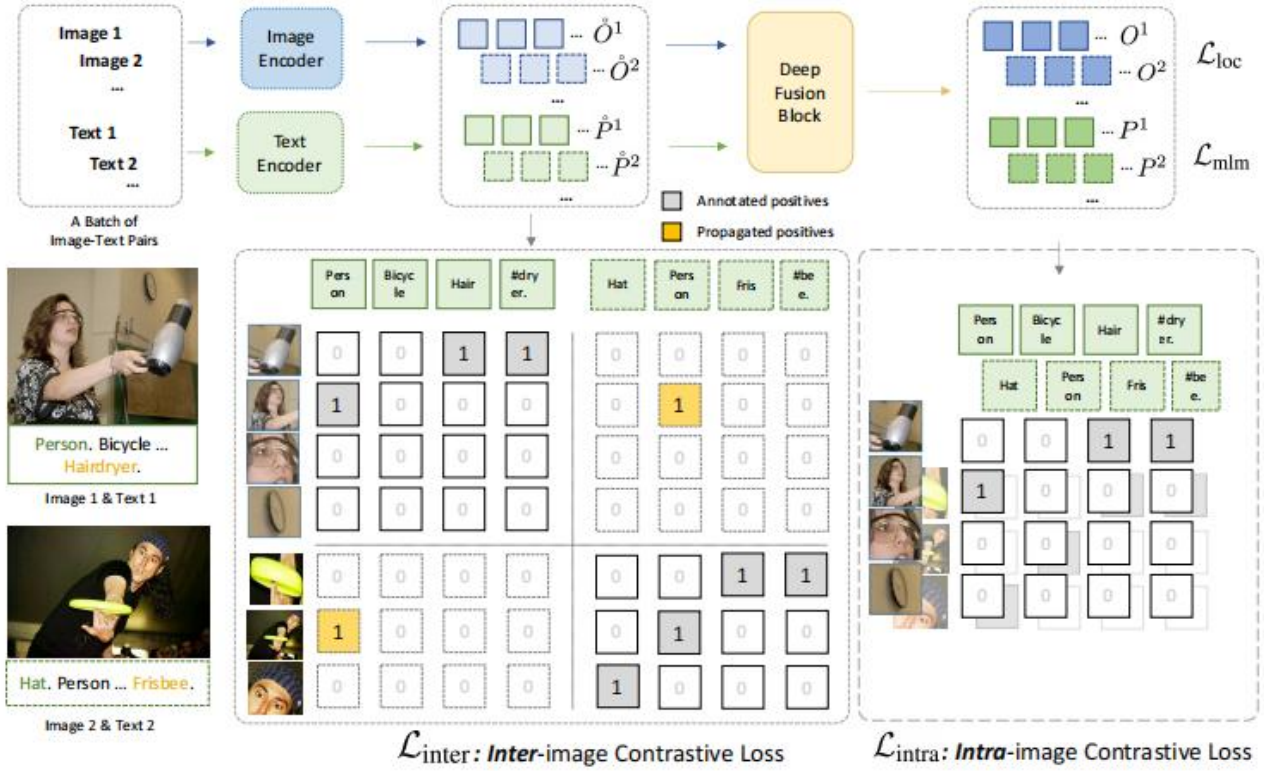
Haotian Zhang, et al. “GLIPv2: Unifying Localization and Vision-Language Understanding.” NeurIPS 2022.

研究现状-知识获取-多任务数据-GLIPv2

GLIPv2: Unifying Localization and VL Understanding

GLIPv2 Pre-training：

- $\mathcal{L}_{\text{GLIPv2}} = \underbrace{\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{intra}}}_{\mathcal{L}_{\text{ground}}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{mlm}}$
- $\mathcal{L}_{\text{mlm}}$ 是BERT中提出的标准masked language modeling loss



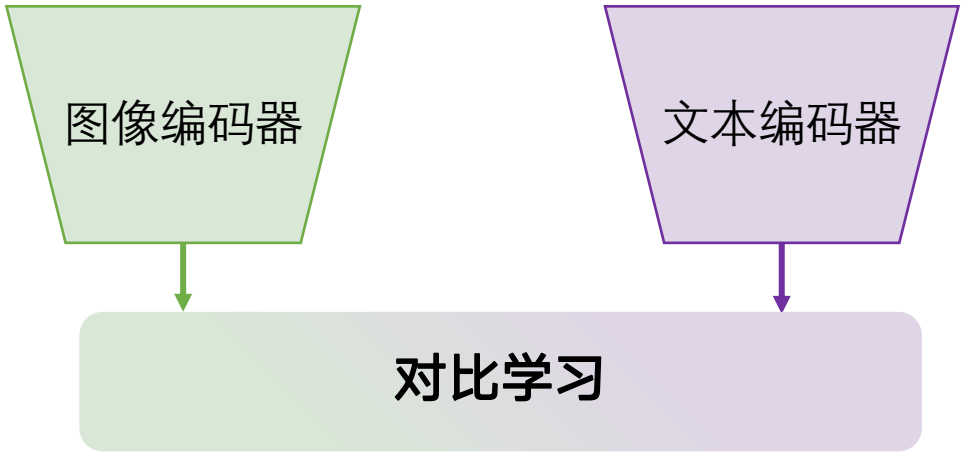
Haotian Zhang, et al. “GLIPv2: Unifying Localization and Vision-Language Understanding.” NeurIPS 2022.

研究现状-知识获取-图像文本对

利用海量图像-文本对，将图像与文本映射到同一嵌入空间，实现概念理解

“a girl is flying a kite”

“a man is riding a motorcycle”



特点：

- **图文关联弱，噪声较大**
- **数据易获取**
- **泛化能力强**

研究现状-知识获取-图像文本对-CLIP

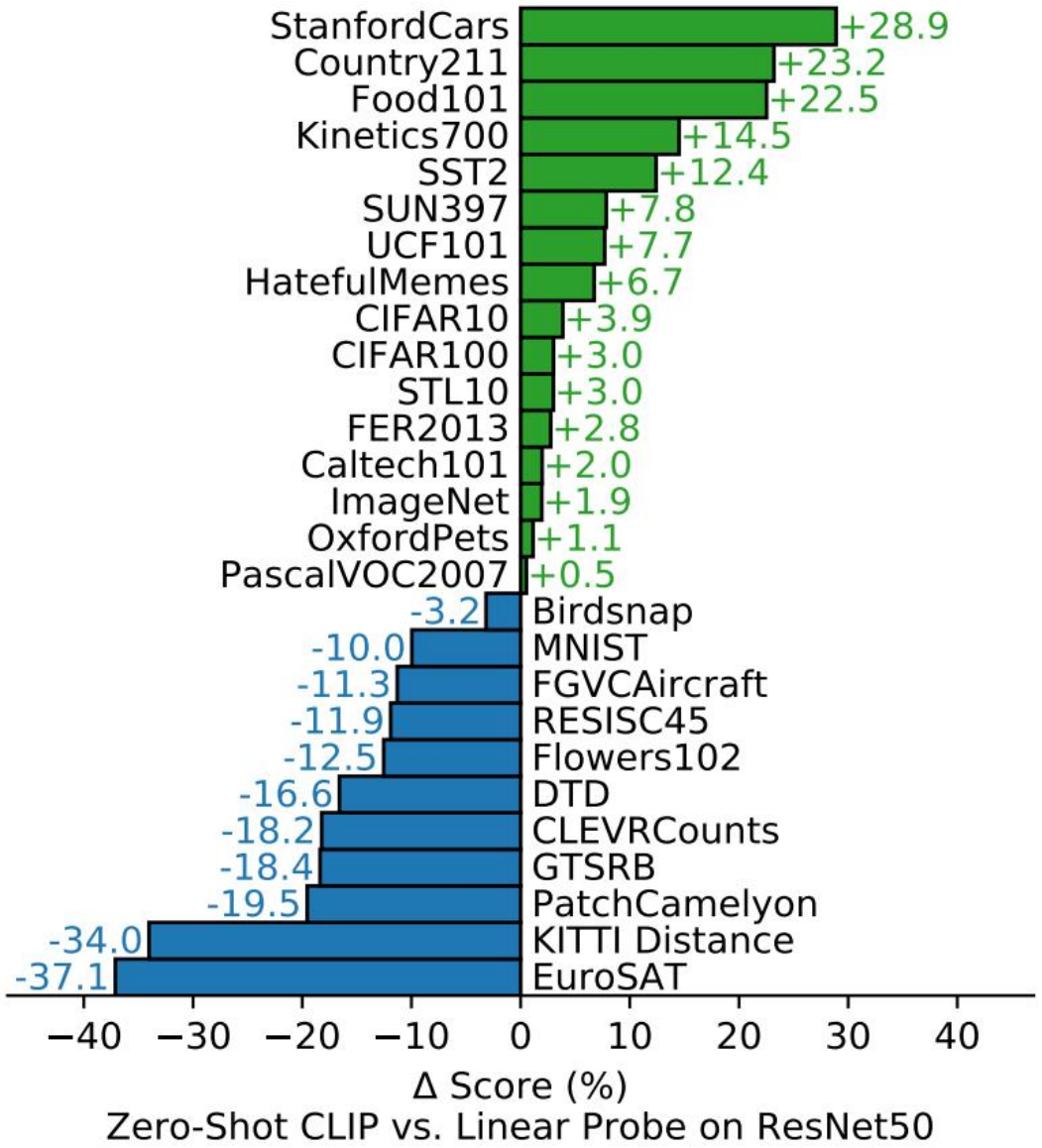
CLIP

首个将Transformer、对比学习结合并提出大型互联网图文对数据集进行训练的模型，拥有较强的泛化能力。

实验发现仅训练一个简单的线性分类头，在大多数数据集上拥有比监督学习的resnet-50有更好的性能。

特点：

- zero-shot能力强
- 需要更大的数据集



Alec Radford, et al. “Learning Transferable Visual Models From Natural Language Supervision.” ICML 2021.

研究现状-知识获取-图像文本对-CLIP

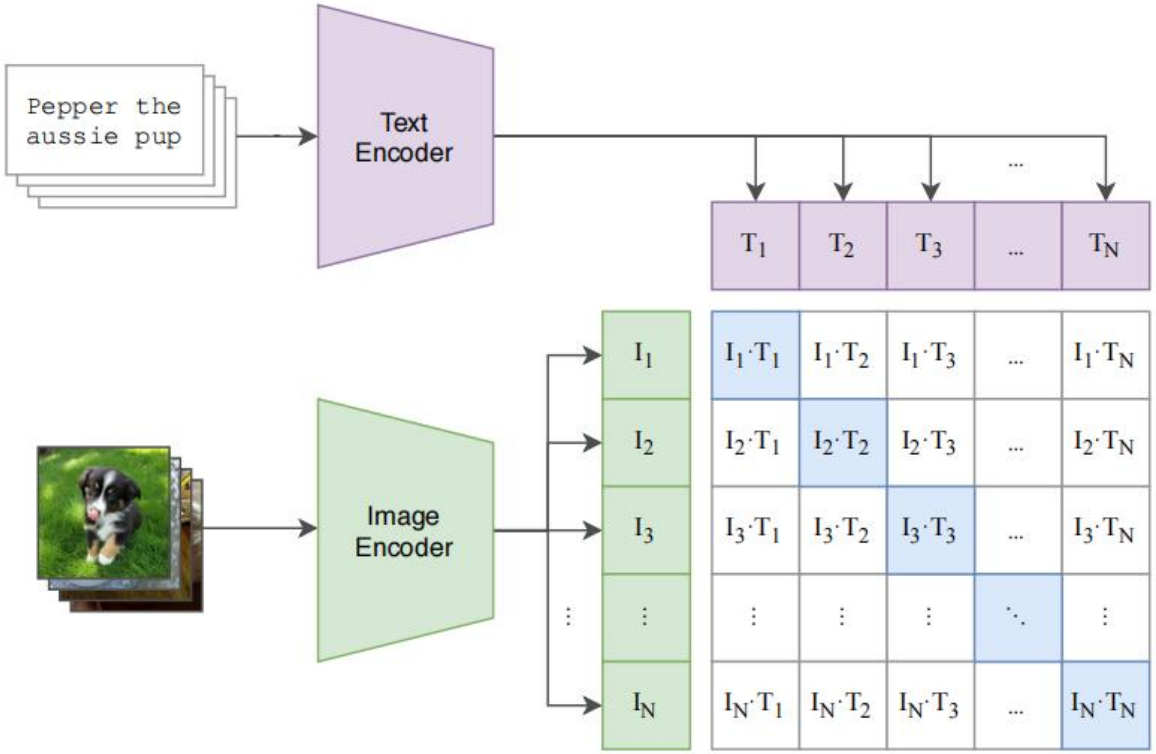
CLIP

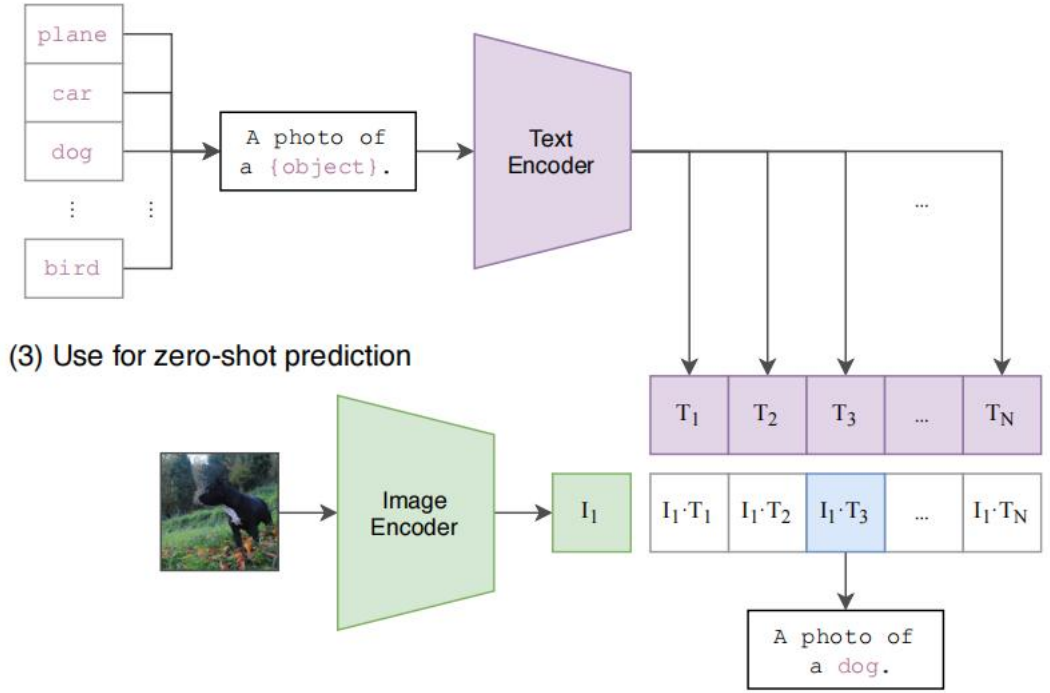
训练：

1. 从数据集中选取大量图文对作为batch
2. 分别经过图文编码器转化成特征
3. 经过线性层投影至同意特征空间。
4. 计算两两图文之间的相似度，并用交叉熵损失函数计算损失。

测试：

1. 通过设定prompt template来处理分类任务，实验发现在分类任务中，使用“a photo of a {object}”对模型性能的提升最为显著
2. 图片通过编码器计算特征后依次和所有设定好的标签prompt计算相似度来进行类型预测





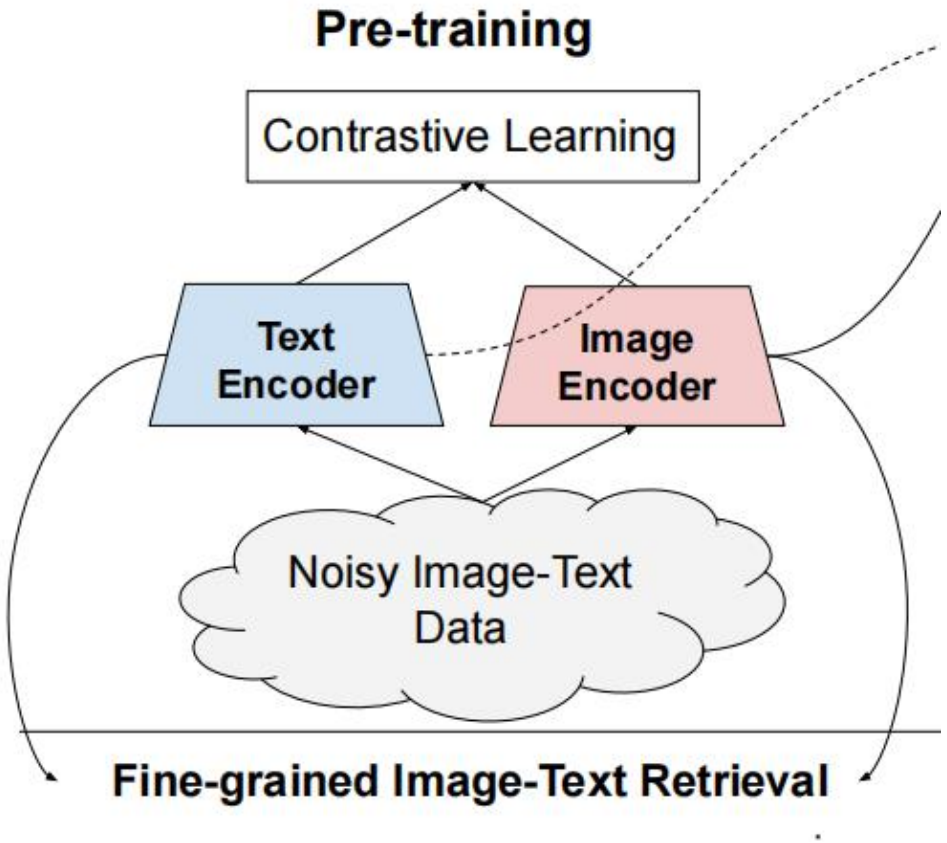
Alec Radford, et al. “Learning Transferable Visual Models From Natural Language Supervision.” ICML 2021.

研究现状-知识获取-图像文本对-ALIGN

CLIP使用的互联网数据集仍然需要对数据集进行清理，这对数据集的扩张带来了不便。本文证明即使不经过数据清理，使用足够大的带有噪声的数据集仍然能够训练出拥有相当性能的模型。

特点：

- 遵循自然图文对数据分布
- 需要更大的数据集
- 通过简单的基于数据频率的过滤来处理数据



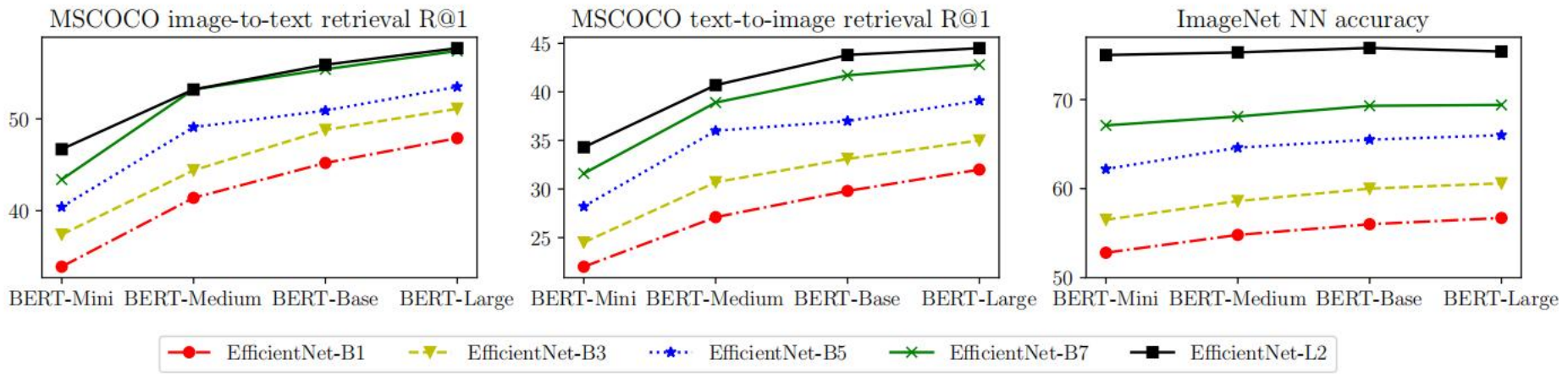
Jia C , Yang Y , Xia Y ,et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision，ICML 2021

研究现状-知识获取-图像文本对-ALIGN

可以看到使用带噪的数据，模型的性能依然可以达到、甚至超越CLIP的性能

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

经过消融实验可以测试不同的特征提取器对模型性能带来的改变。一般而言，更大的模型会具有更高的精度。



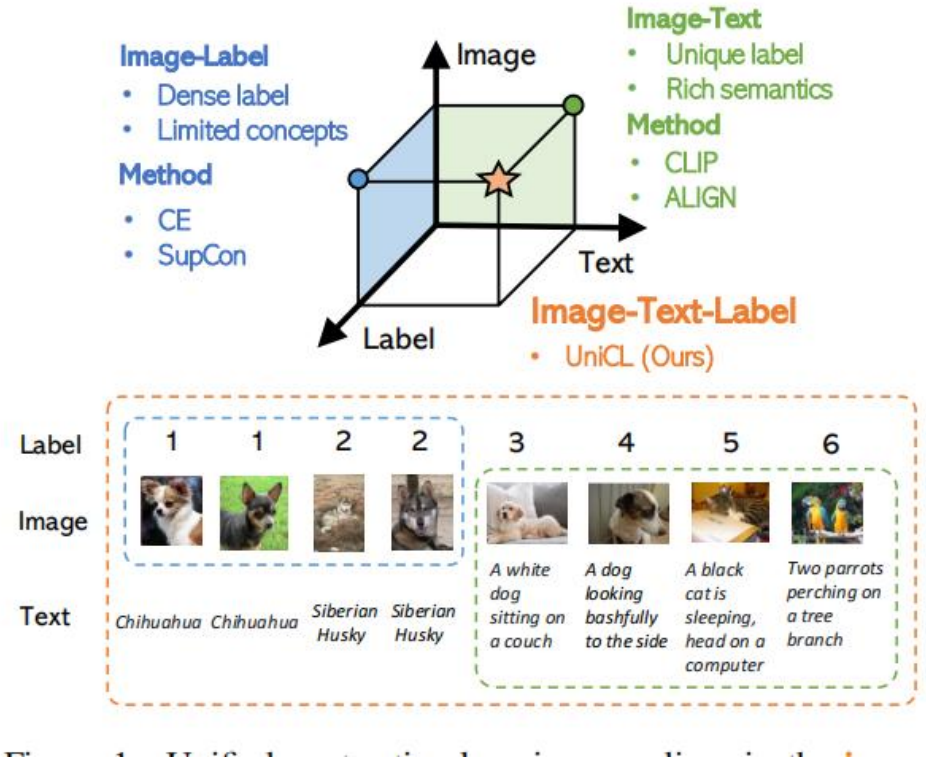
Jia C , Yang Y , Xia Y ,et al.Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

研究现状-知识获取-图像文本对-UniCL

对比学习作为一种**无监督学习方法**，得到预训练模型具有**更强的泛化能力**，然而在**断能力方面会有所下降。**

特点：

- 构建了**图片-文本-标签**的新对比学习范
- 使用了监督信号来加强模型的判别能力

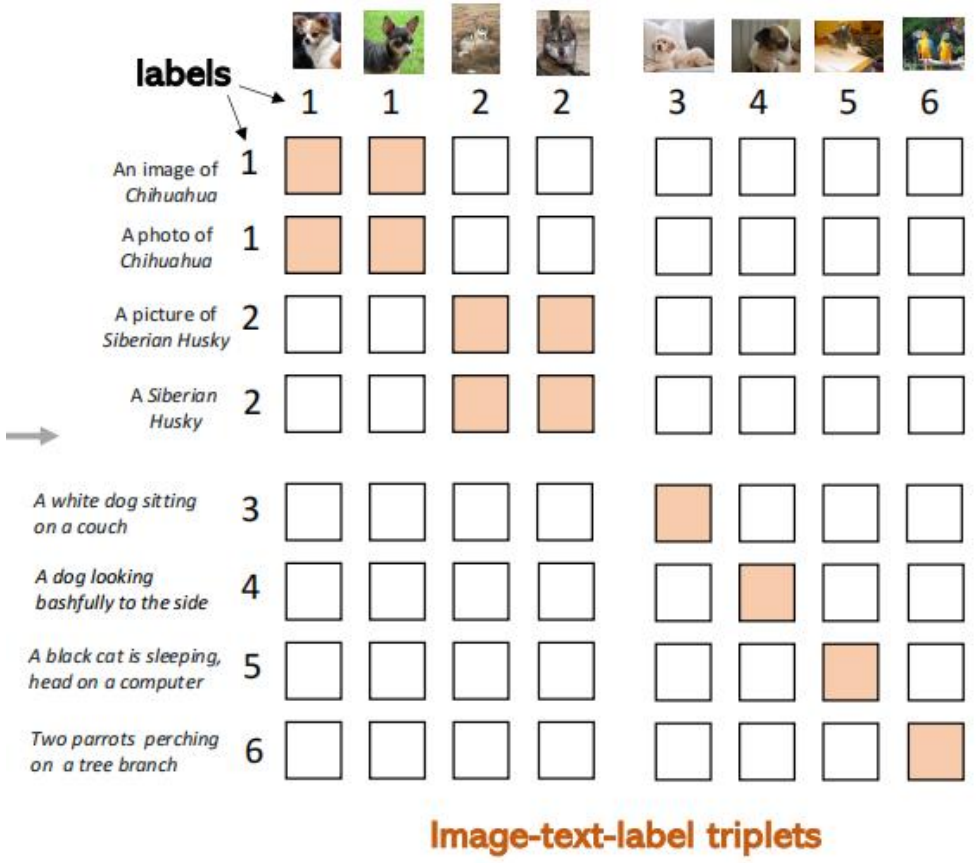


Yang J, Li C, et al. Unified contrastive learning in image-text-label space, CVPR 2022.

研究现状-知识获取-图像文本对-UniCL

论文将图片-文本-标签对视为**具有标签的图文对**。

简单来说，在一个batch内，可能具有相同标签的图片，因此需要对损失函数进行修改来考虑监督信号。



Yang J, Li C, et al. Unified contrastive learning in image-text-label space, CVPR 2022.

研究现状-知识获取-图像文本对-UniCL

可以从损失函数中看出UniCL与CLIP的区别，即在一个batch内存在一对多的情况时，需要同时拉近多个目标的相似度——当仅考虑文本信息，函数退化为与CLIP一致

- The image-to-text contrastive loss to align matched images in a batch with a given text

$$\mathcal{L}_{i2t} = -\sum_{i\in\mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_k)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j)} \quad (2)$$

where $k \in \mathcal{P}(i) = \{k|k \in \mathcal{B}, y_k = y_i\}$.

- The text-to-image contrastive loss to align matched texts to a given image

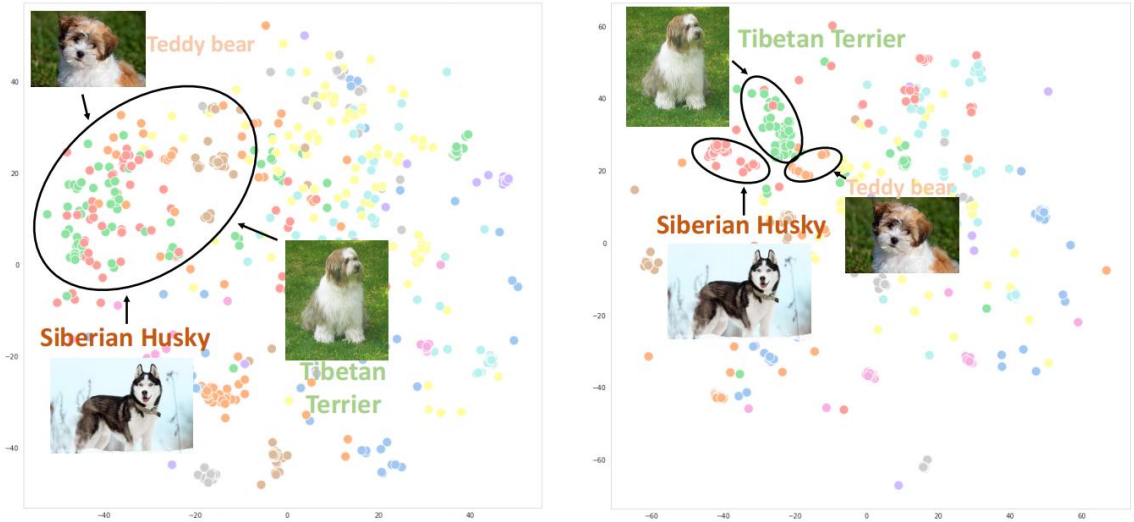
$$\mathcal{L}_{t2i} = -\sum_{j\in\mathcal{B}} \frac{1}{|\mathcal{P}(j)|} \sum_{k\in\mathcal{P}(j)} \log \frac{\exp(\tau \boldsymbol{u}_k^T \boldsymbol{v}_j)}{\sum_{i\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j)} \quad (3)$$

where $k \in \mathcal{P}(j) = \{k|k \in \mathcal{B}, y_k = y_j\}$.

Yang J, Li C, et al. Unified contrastive learning in image-text-label space, CVPR 2022.

研究现状-知识获取-图像文本对-UniCL

从可视化结果中可以看到UniCL在对数据的判别能力上有了提升



UniCL加强了文本与标签之间的联系，使得模型具有更加强大的判别能力

Training Data	Method	Metric			
		Zero-shot		ImageNet-1K Finetuning	Linear Probe 18 datasets
		ImageNet-1K	14 datasets		
YFCC-14M	CLIP	30.1	36.3	77.5	72.7
ImageNet-21K	UniCL	28.5	37.8	78.8	80.5
YFCC-14M(half) + ImageNet-21K(half)	UniCL	36.4	45.5	79.0	80.0
YFCC-14M(half) + ImageNet-21K(half)	Multi-task	33.0	41.5	78.0	74.1
YFCC-14M + ImageNet-21K	UniCL	**40.5**	**49.1**	**80.2**	**81.6**
ImageNet-22K	UniCL	66.8	38.9	80.3	82.0
YFCC-14M + ImageNet-22K	UniCL	70.5	52.4	**80.5**	82.0
YFCC-14M + ImageNet-22K	Multi-task	40.9	47.6	80.4	82.0
GCC-15M + ImageNet-22K	UniCL	**71.3**	**53.8**	80.0	82.1
GCC-15M + ImageNet-22K	Multi-task	50.6	51.8	79.9	**82.5**

Yang J, Li C, et al. Unified contrastive learning in image-text-label space, CVPR 2022.

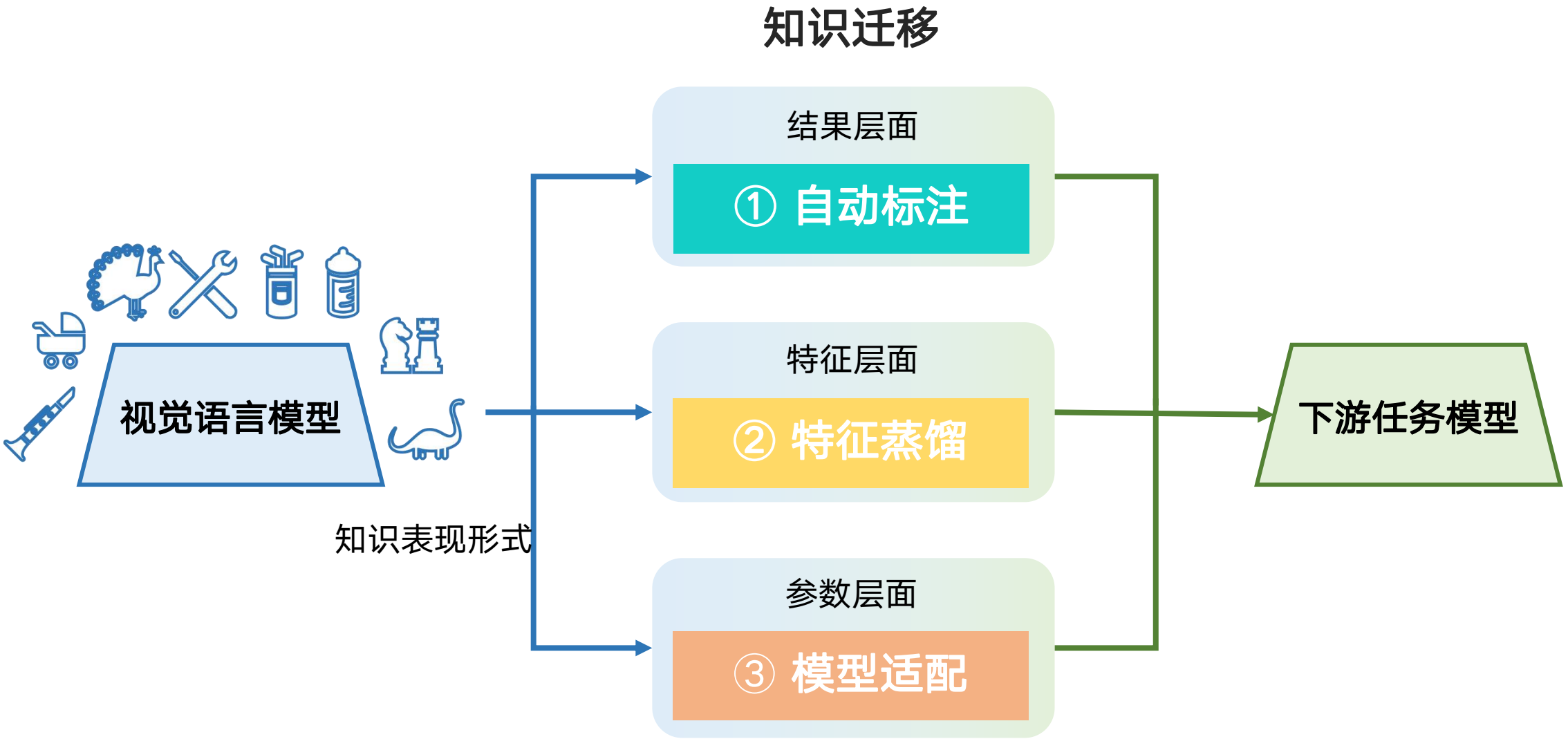
研究现状-知识获取

视觉语言模型具有开放感知能力，但无法执行具体的下游任务，需要进行知识迁移

	CLIP	Mask R-CNN	CLIP+ Mask R-CNN
训练数据量	400M	100K	100K
训练资源	3072 GPU天	16 GPU天	32 GPU天
开放感知能力	✓	✗	✓
目标检测能力	✗	✓	✓

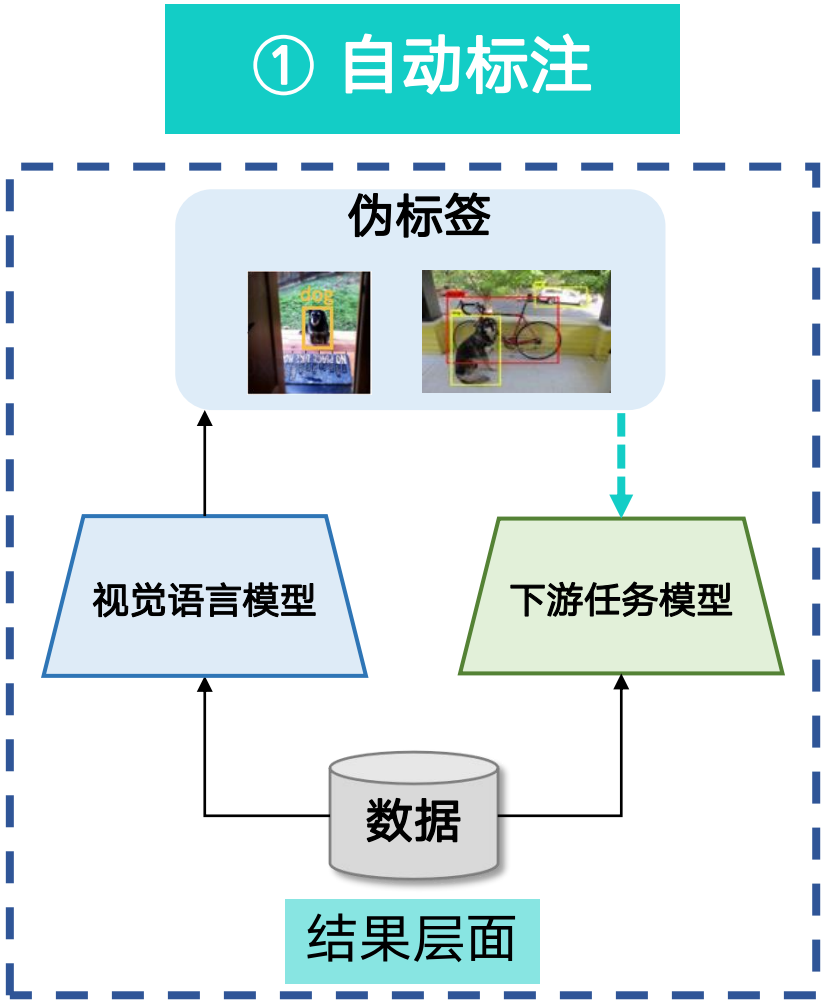
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. “Learning Transferable Visual Models From Natural Language Supervision.” ICML 2021.
Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. “ Mask R-CNN.” ICCV 2017.

研究现状-知识获取



研究现状-知识获取

利用视觉语言模型生成下游任务的**伪标签**，用于下游任务模型训练



特点：

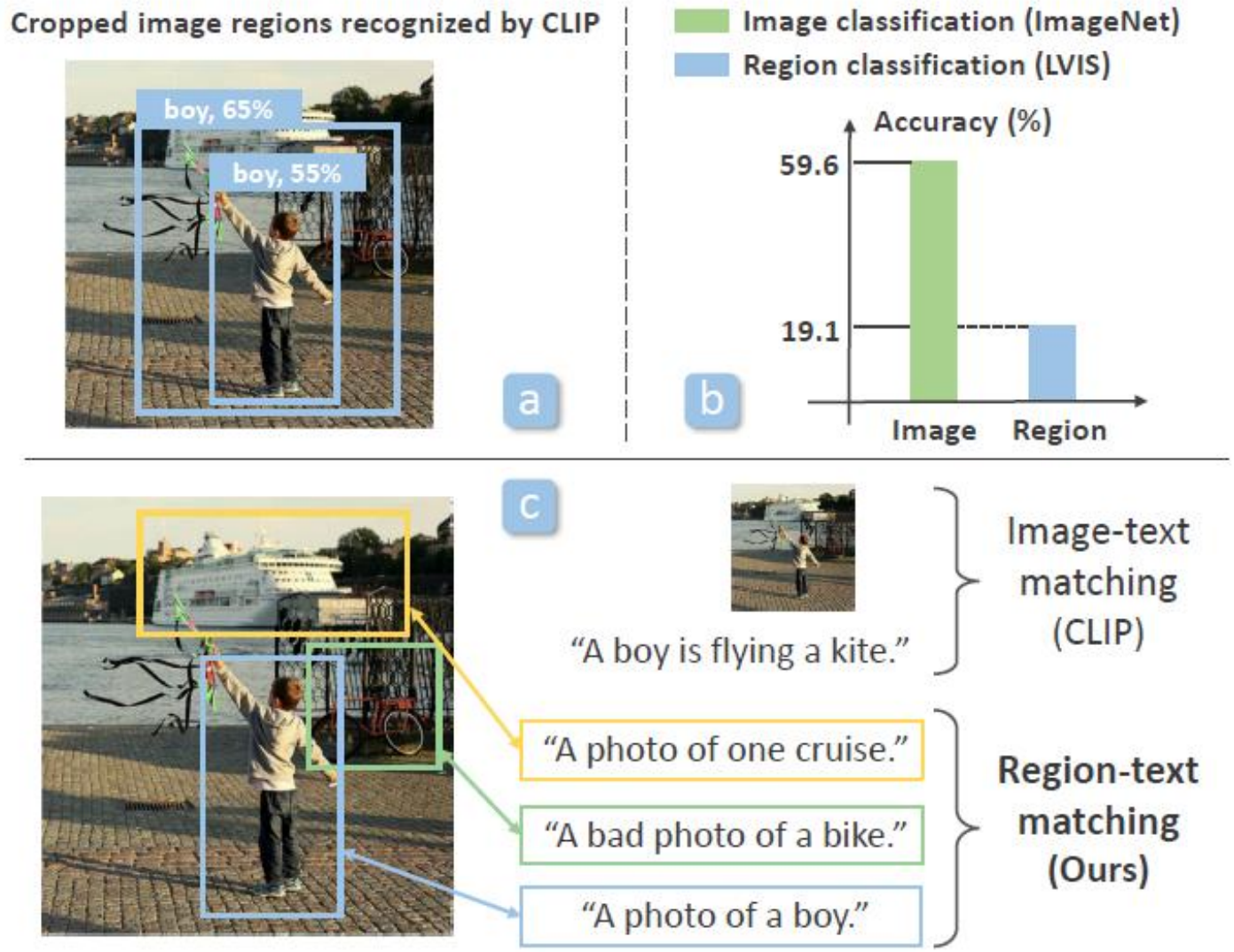
- **结果层**知识迁移
- 人工设计**标注策略**
- **伪标签质量**影响迁移效果

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, et al. “RegionCLIP: Region-based Language-Image Pretraining.” CVPR 2022.

研究现状-知识获取-自动标注-RegionCLIP

CLIP在识别图片整体时表现较好，但在识别图像局部时表象欠佳

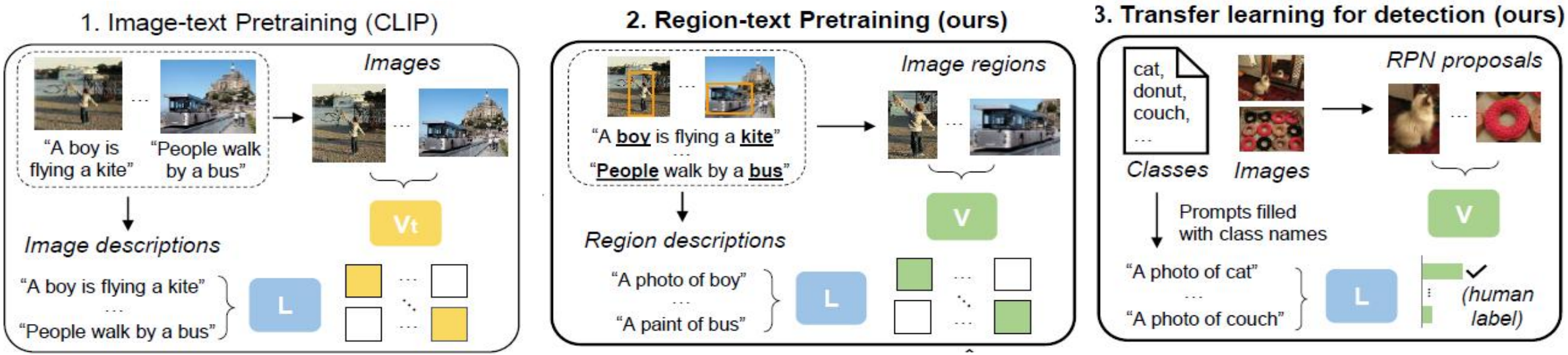
RegionCLIP：利用图像局部以及其对应的伪标签生成的数据集训练模型，提高区域识别能力



Yiwu Zhong, et al. "Exploiting Unlabeled Data with Vision and Language Models for Object Detection." CVPR 2022.

研究现状-知识获取-自动标注-RegionCLIP

RegionCLIP训练方法



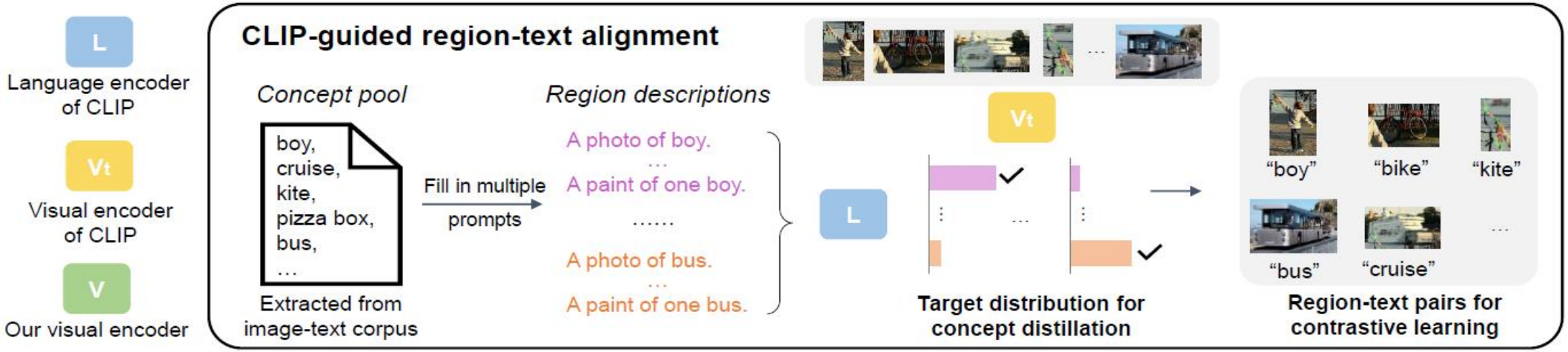
- 使用训练好的视觉语言模型，实现图片与文本的匹配
- 用RPN获得有意义的区域，然后再将这些区域与的文本描述相匹配。
- 使用区域级别的图像-文本数据集对模型进行训练，训练好的模型可以有效对局部图像生成文本描述

Yiwu Zhong, et al. "Exploiting Unlabeled Data with Vision and Language Models for Object Detection." CVPR 2022.

研究现状-知识获取-自动标注-RegionCLIP

生成伪标签，将区域和文本匹配

- 为每一个概念池中的概念创建一个简单的句子，使用文本编码器对其进行编码，就得到了所有图像可能对应的文本，$\{l_i\}$。
- 通过CLIP得到区域图像的特征v_i(可能是不完善的），再分别计算v_i与 $\{l_i\}$ 的 匹配分数，取拥有最高分数文本，记为 l_m，就得到了每一个区域图像和对应伪标签的组合 $\{v_i, l_m\}$



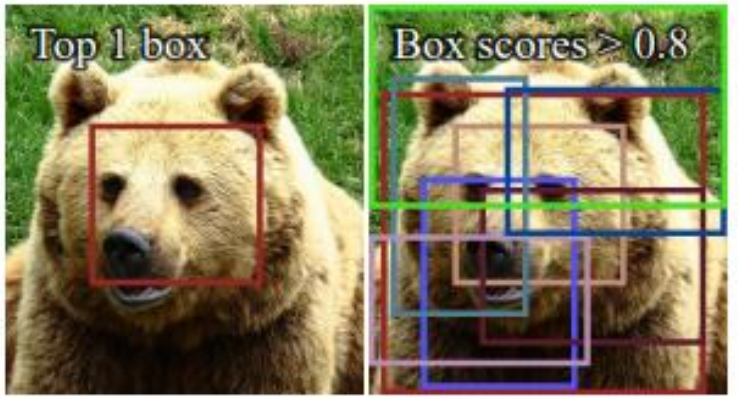
Yiwu Zhong, et al. "Exploiting Unlabeled Data with Vision and Language Models for Object Detection." CVPR 2022.

研究现状-知识获取-自动标注-VL PLM

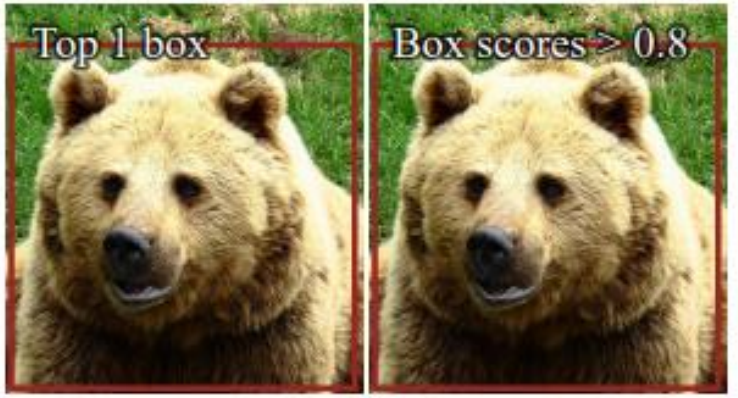
VL-PLM：基于视觉语言模型的伪标签生成

视觉语言模型：可以通过网络抓取数据，在不需要人工注释的情况下获得图像-文本对数据集，进行训练。

- 视觉语言模型：目标定位质量较低
- 两阶段类别无关的区域生成网络：辅助提高目标定位质量



CLIP on raw region proposals



CLIP on enhanced region proposals

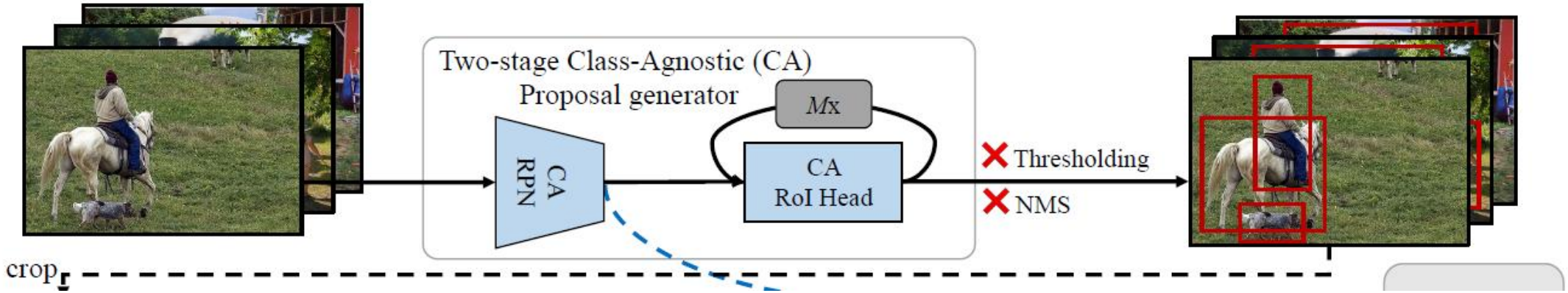
Shiyu Zhao, et al. “Exploiting Unlabeled Data with Vision and Language Models for Object Detection.” ECCV 2022.

研究现状-知识获取-自动标注-VL PLM

两阶段类别无关的区域生成网络

区域生成器不仅能够定位训练期间看到的类别对象，还应该能够定位**新类别**的对象。

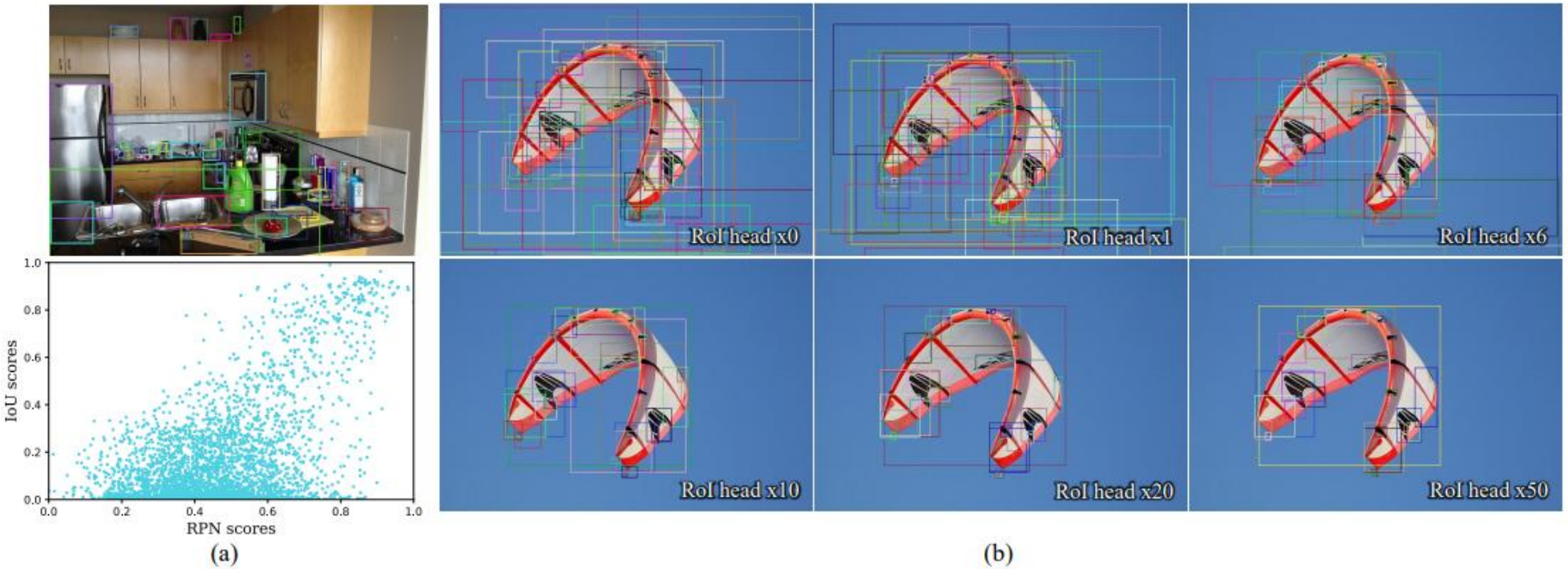
- 两阶段 RPN对新类别有很好的泛化能力
- 忽略了训练集的类别信息，训练类别无关的网络，进一步提高泛化能力



Shiyu Zhao, et al. “Exploiting Unlabeled Data with Vision and Language Models for Object Detection.” ECCV 2022.

研究现状-知识获取-自动标注-VL PLM

- 使用RPN模型对区域进行预测：RPN分数与预测区域和真实区域的IoU分数成正相关，是评判区域定位质量的很好的指标。
- RoI头部可以改善定位质量：将预测区域多次输入到RoI头部，从而找到更好的定位边界框，提供更好的伪标签

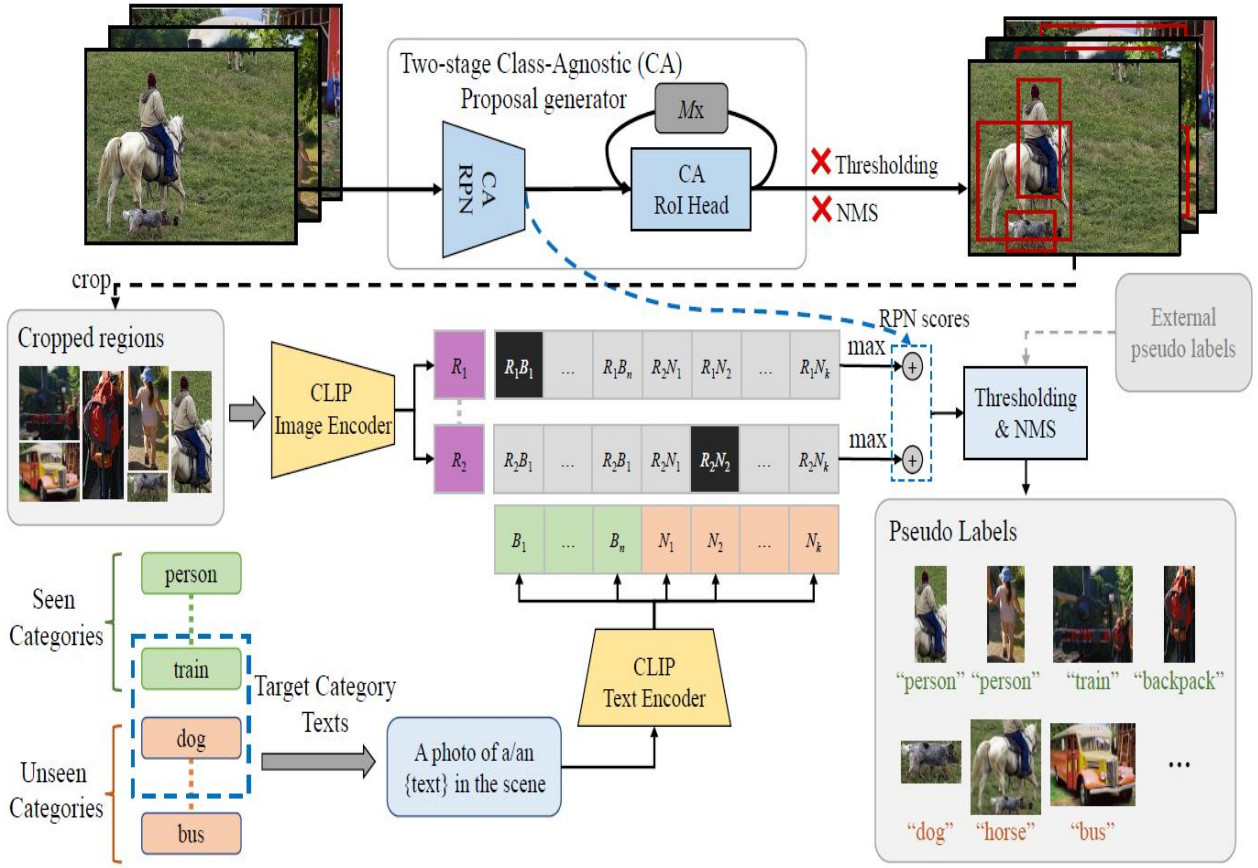


Shiyu Zhao, et al. “Exploiting Unlabeled Data with Vision and Language Models for Object Detection.” ECCV 2022.

研究现状-知识获取-自动标注-VL PLM

伪标签生成过程

- 将未标记的图像输入两阶段类别无关检测器以**获得区域图像**
- 将区域图像输入到CLIP图像编码器中，获得在CLIP空间中的**特征**。
- 计算每个区域的图像特征和文本特征之间的**相似性**，综合利用**类别无关检测器**和**视觉语言模型**的评估分数，得到最终的伪标签



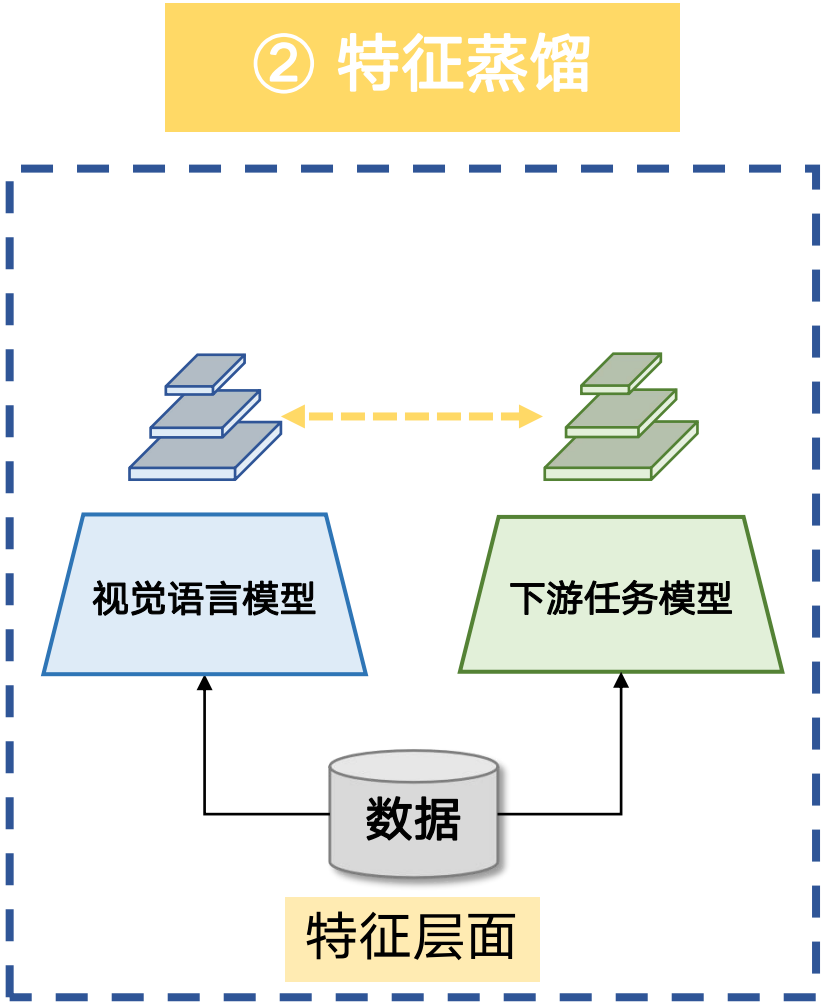
Shiyu Zhao, et al. "Exploiting Unlabeled Data with Vision and Language Models for Object Detection." ECCV 2022.

研究现状-知识获取

拉近下游任务模型和视觉语言模型**特征**之间的距离，迁移深度知识

特点：

- **特征层**知识迁移
- 人工设计**特征对齐方法**
- 获得语言对齐的视觉特征



Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, Yin Cui. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.” ICLR 2022.

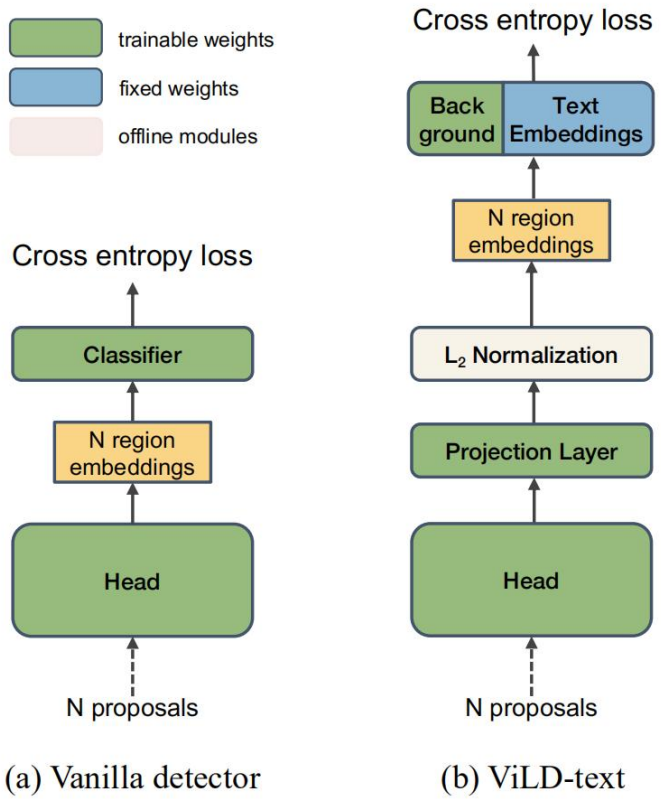
研究现状-知识获取-特征蒸馏-ViLD

ViLD（Vision and Language knowledge Distillation）

开放词汇对象检测

(Openvocabulary object detection, OVD)

- Vanilla：传统的两阶段检测器（例如Mask R-CNN）的分类部分。
- ViLD-text：通过将类型名称输入预训练的文本编码器(CLIP)来获得文本特征。通过使用文本特征向量进行分类来训练区域特征向量。



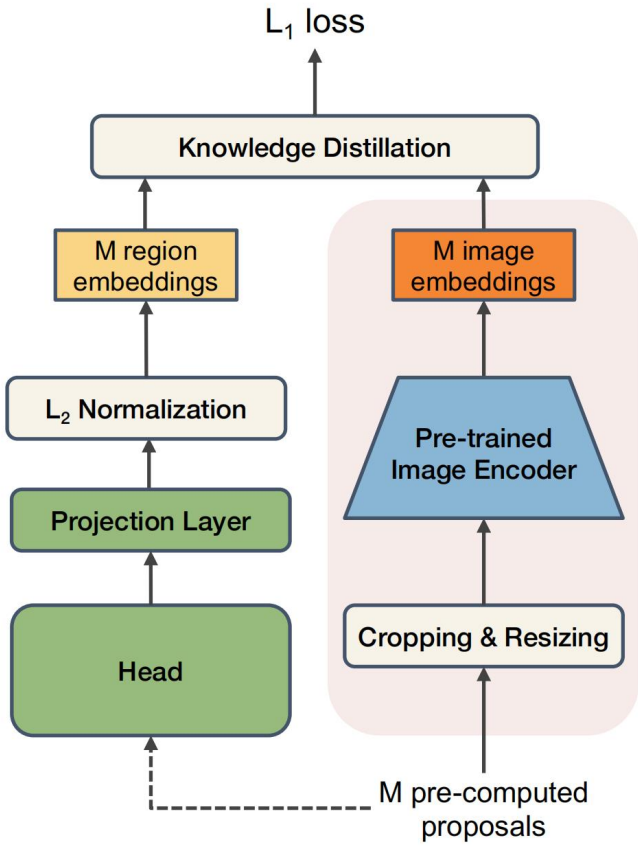
Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo and Yin Cui. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.” ICLR 2022.

研究现状-知识获取-特征蒸馏-ViLD

开放词汇对象检测

(Openvocabulary object detection, OVD)

- ViLD-image：算法首先从预训练的文本编码器(CLIP)中获取图像特征，然后使用这些图像特征与检测框的区域特征进行对齐。和ViLD-Text处理的不同是，ViLD-image同时从基础类别和新的类别中做了知识蒸馏。

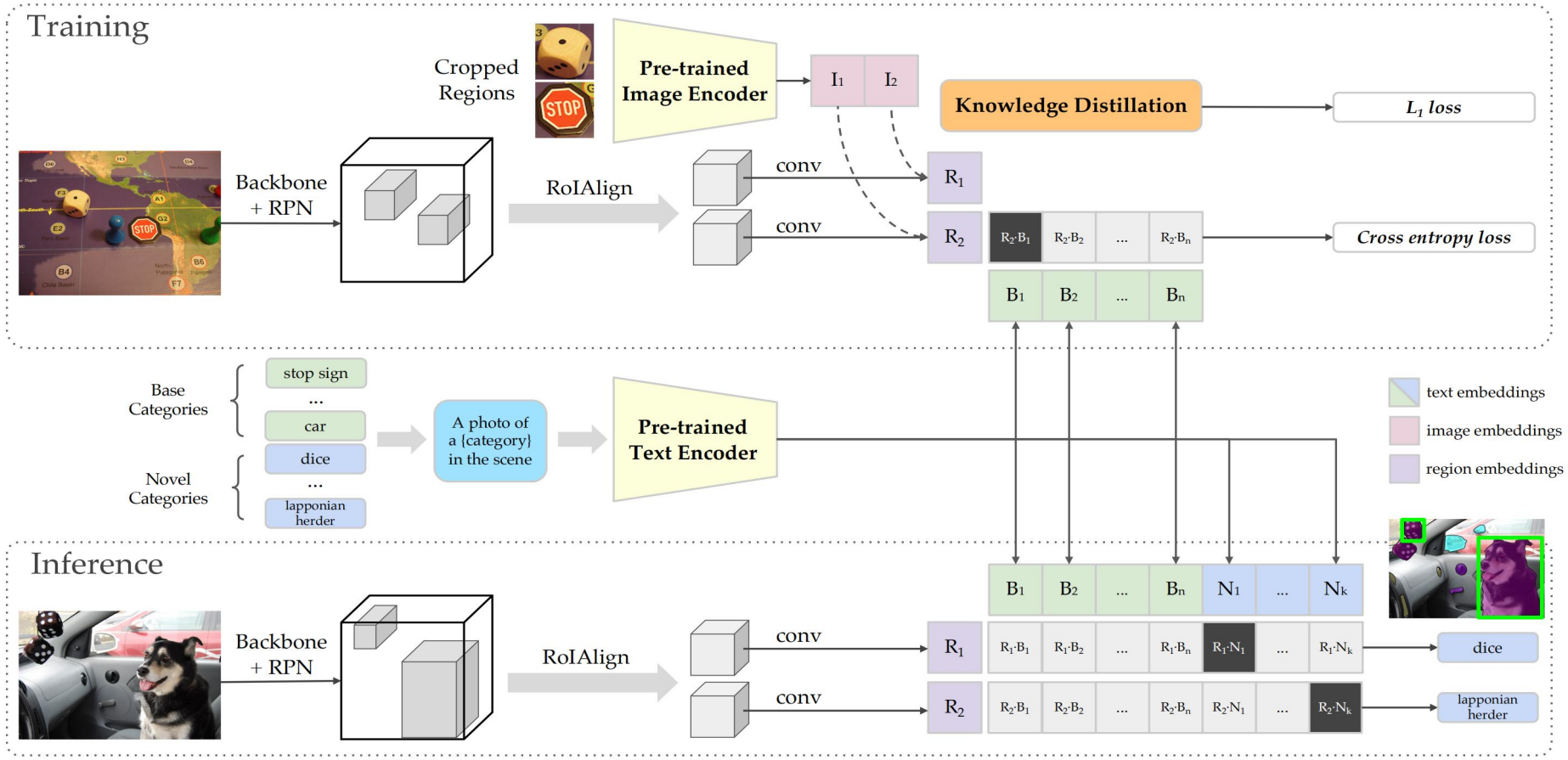


(c) ViLD-image

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo and Yin Cui. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.” ICLR 2022.

研究现状-知识获取-特征蒸馏-ViLD

- Vild由ViLD-Text和ViLD-image两部分构成。
- 从预训练的开放词汇图像分类多模态模型（例如CLIP）中**蒸馏区域级知识**。



Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo and Yin Cui. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.” ICLR 2022.

研究现状-知识获取-特征蒸馏-ViLD

实验结果

- 使用CLIP进行开放词汇检测在新颖类别上的性能远超过了监督学习方法。
- ViLD在LVIS数据集上取得了先进性能。

Method	AP_r	AP_c	AP_f	AP
Supervised (base class only)	0.0	22.6	32.4	22.5
CLIP on cropped regions w/o objectness	13.0	10.6	6.0	9.2
CLIP on cropped regions	**18.9**	18.8	16.0	17.7
Supervised (base+novel)	4.1	23.5	33.2	23.9
Supervised-RFS (base+novel)	12.3	24.3	32.4	25.4

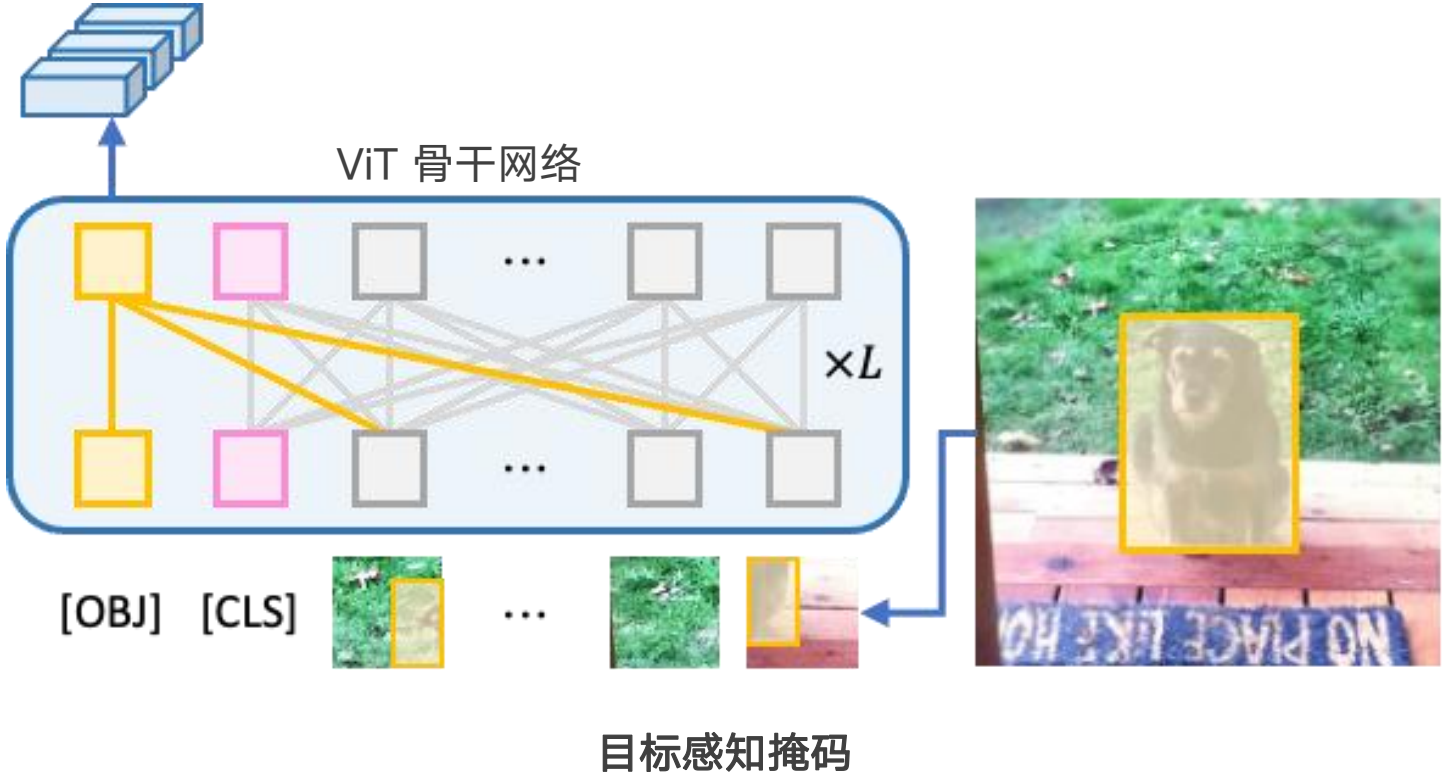
Backbone	Method	AP_r	AP_c	AP_f	AP
ResNet-50+ViT-B/32	CLIP on cropped regions†	18.9	18.8	16.0	17.7
	ViLD-text+CLIP†	**22.6**	24.8	29.2	26.1
ResNet-50	Supervised-RFS (base+novel)	12.3	24.3	32.4	25.4
	GloVe baseline	3.0	20.1	30.4	21.2
	ViLD-text	10.1	23.9	32.5	24.9
	ViLD-image	11.2	11.3	11.1	11.2
	ViLD (w=0.5)	16.1	20.0	28.3	22.5
	ViLD-ensemble (w=0.5)	**16.6**	24.6	30.3	25.5
EfficientNet-b7	ViLD-ensemble w/ ViT-L/14 (w=1.0)	21.7	29.1	33.6	29.6
	ViLD-ensemble w/ ALIGN (w=1.0)	**26.3**	27.2	32.9	29.3
ResNeSt269+HTC	2020 Challenge winner (Tan et al., 2020)‡	30.0	41.9	46.0	41.5

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo and Yin Cui. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.” ICLR 2022.

研究现状-知识获取-特征蒸馏-OADP

OADP：设计**目标感知掩码**针对各目标提取针对性深度知识，通过**金字塔知识蒸馏**迁移至目标检测器，并与预训练文本编码器提取的**类别深度知识**相匹配，从而对未知类别目标精准分类

- 自适应与掩码：引入目标区域自适应变形以保证**知识完整性**，结合目标感知掩码过滤噪声。

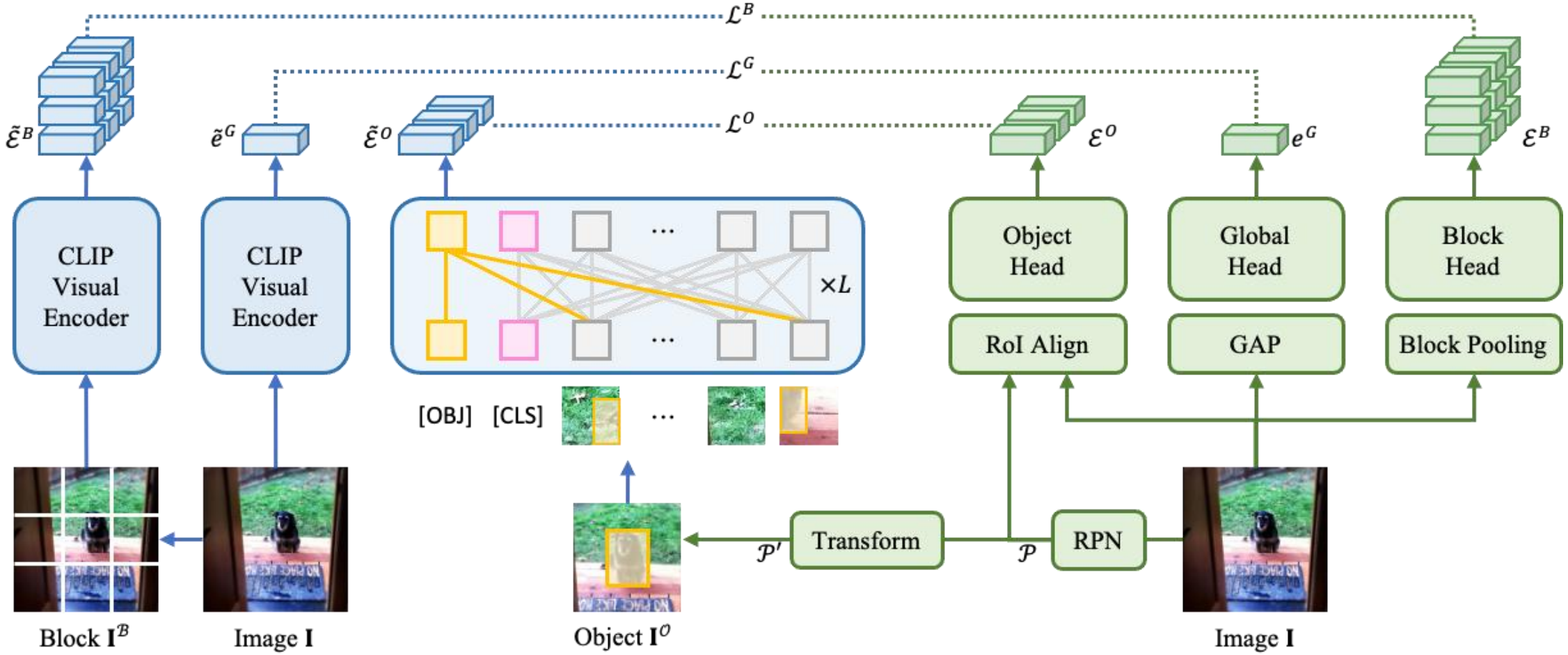


Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, Si Liu. “Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection.” CVPR 2023.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, Si Liu. “GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection.” CVPR 2022.

研究现状-知识获取-特征蒸馏-OADP

- 金字塔知识蒸馏：通过知识蒸馏，将 CLIP 的多种预训练深度视觉知识(全局、区块、目标)迁移至目标检测器，使其具备识别未知类别目标并**提取有效视觉知识**的能力。

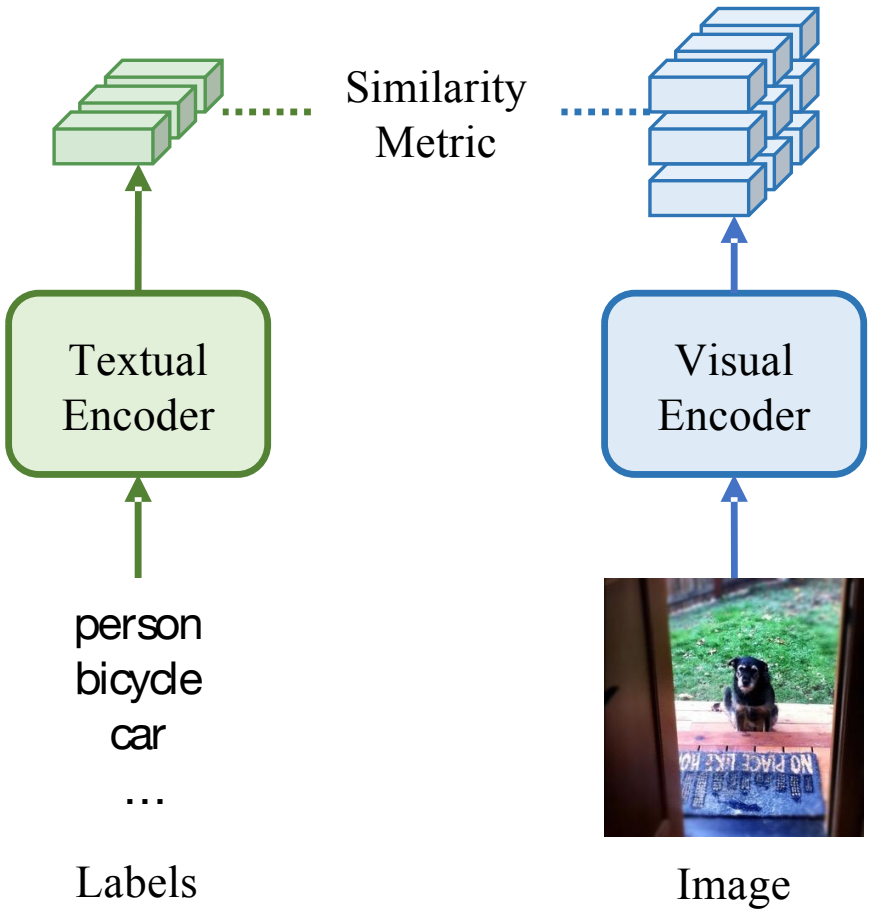


Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, Si Liu. “Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection.” CVPR 2023.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, Si Liu. “GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection.” CVPR 2022.

研究现状-知识获取-特征蒸馏-OADP

- 类别知识：与传统封闭集检测算法相比，开放词表检测算法的类别知识不再由数据集给出，而是使用CLIP 文本编码器提取类别文本标签的语义知识，语义信息丰富度更高且泛化性更强。
- 开放词表分类：由于 CLIP 使用统一特征空间表示文本和视觉深度知识，因此可直接计算目标视觉知识与各类别知识的相似程度，将相似度最高的类别作为目标分类预测。

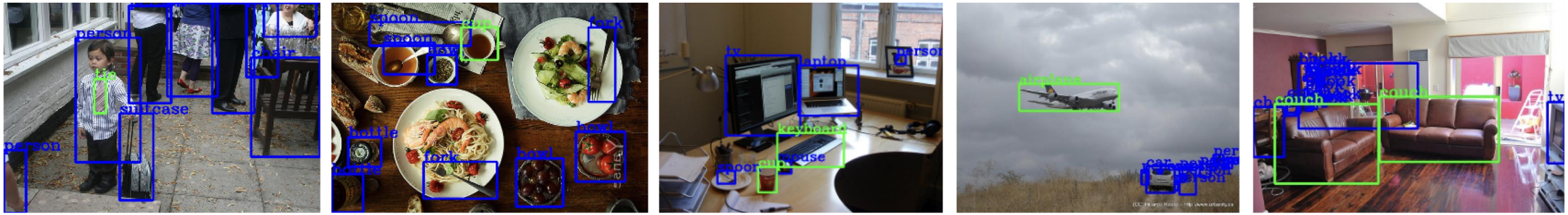


Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, Si Liu. “Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection.” CVPR 2023.
Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, Si Liu. “GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection.” CVPR 2022.

研究现状-知识获取-特征蒸馏-OADP

实验结果

- 开放词表分类可视化：使用 CLIP 对提议框进行开放词表分类，绿色框表示未知类别目标



- 检测器特征可视化：蒸馏后的检测器具备为未知类别目标提取丰富视觉知识的能力

Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, Si Liu. “Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection.” CVPR 2023.

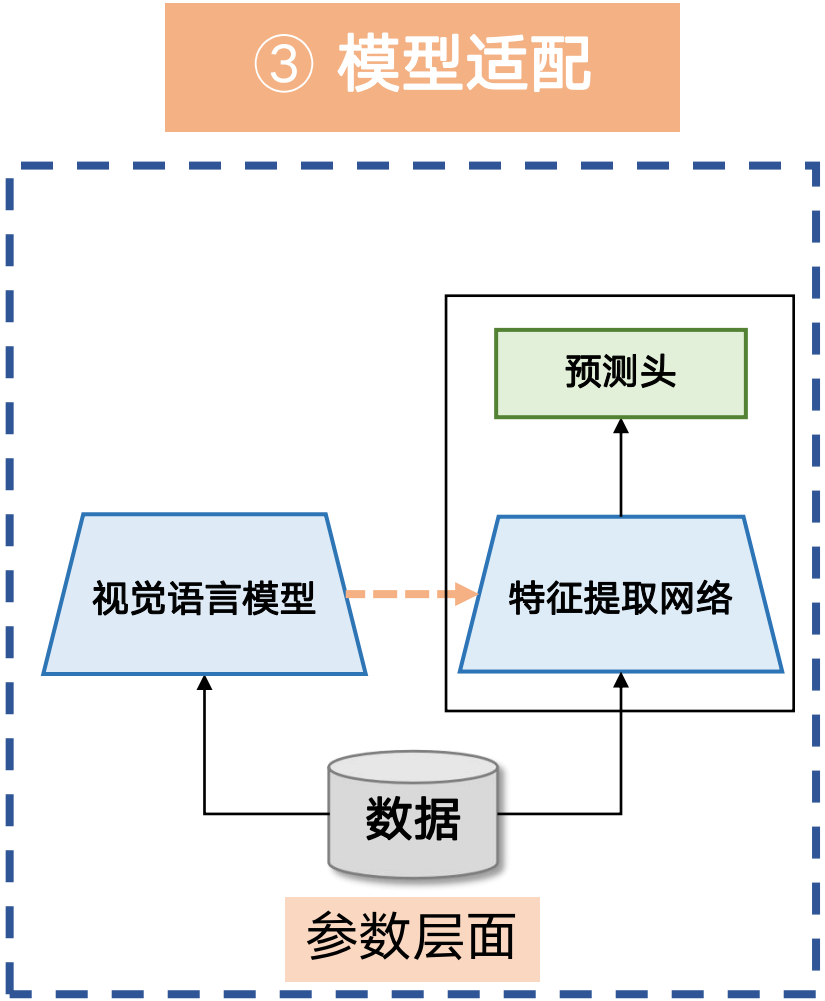
Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, Si Liu. “GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection.” CVPR 2022.

研究现状-知识获取

继承视觉语言模型的权重，引入可学习的下游任务**适配**模块

特点：

- **参数层**知识继承
- 人工设计**适配模块**
- 任务差异影响适配难度



研究现状-知识获取-模型适配-SAN

SAN：通过侧适配网络生成掩码候选和注意力倾向，

将 CLIP 模型应用于开放词表语义分割

动机：

- 大规模视觉语言模型（例如CLIP）使得在开放词汇图像分类成为可能
- 探索如何在语义分割中应用这种方法

挑战：

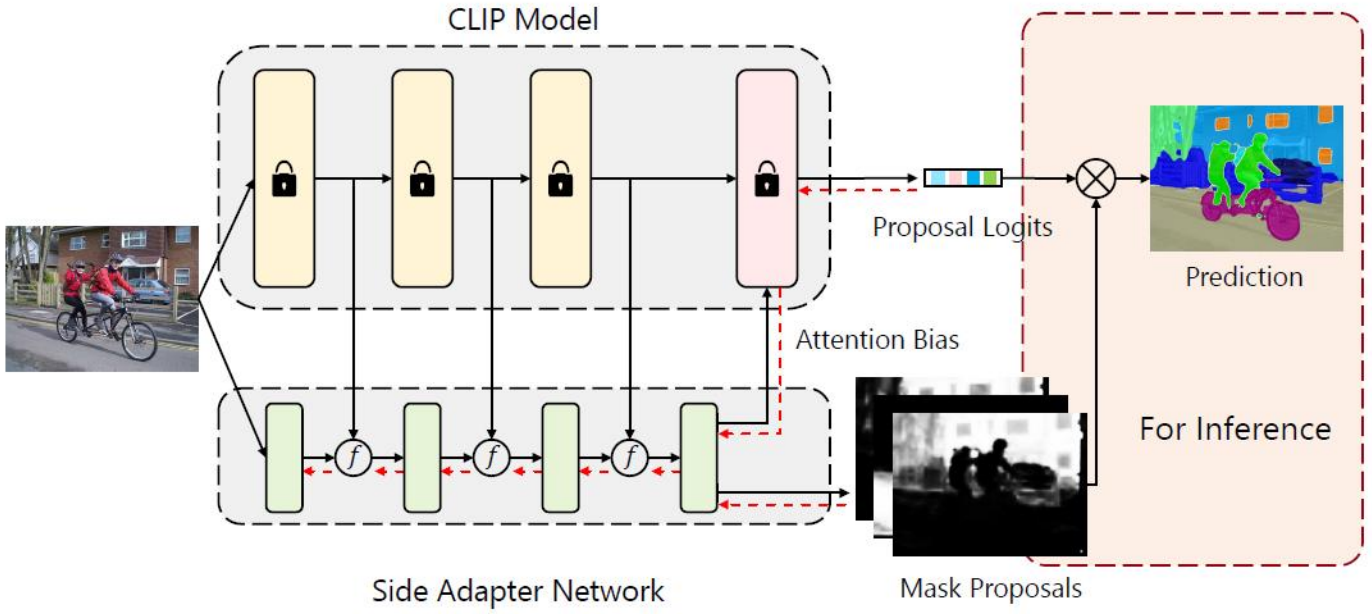
- CLIP 模型在图像级进行训练，缺乏像素级识别能力
- 分割数据集远小于视觉语言预训练数据集
- 分割任务输入图像分辨率远大于 CLIP 可处理范围

Xu, Mengde, et al. "Side adapter network for open-vocabulary semantic segmentation." CVPR 2023.

研究现状-知识获取-模型适配-SAN

方法-总体架构：

- 冻结的CLIP模型做为分类器
- 侧适配网络 SAN 生成掩码候选和注意力倾向，从而指导CLIP模型深层预测相应区域的分类分数
- 在测试过程中，掩码和掩码的分类分数通过矩阵相乘进行组合，以获得最终的预测结果

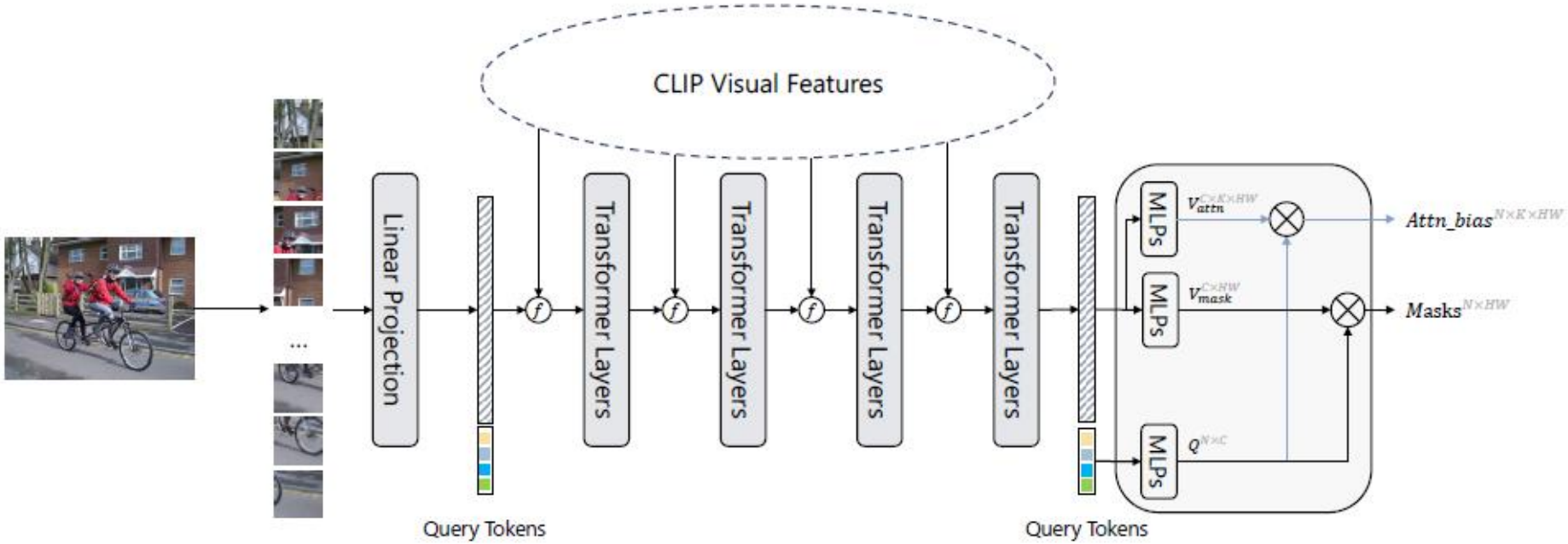


Xu, Mengde, et al. "Side adapter network for open-vocabulary semantic segmentation." CVPR 2023.

研究现状-知识获取-模型适配-SAN

方法-侧适配网络：

- 每个Transformer层中间融合了CLIP模型的即时视觉特征
- 查询和视觉特征使用多层感知机进行编码，以生成注意力倾向和掩码候选
- 注意力偏差引导 CLIP 实现精准的分割图识别，最终通过分割图和分割图类别预测进行语义分割

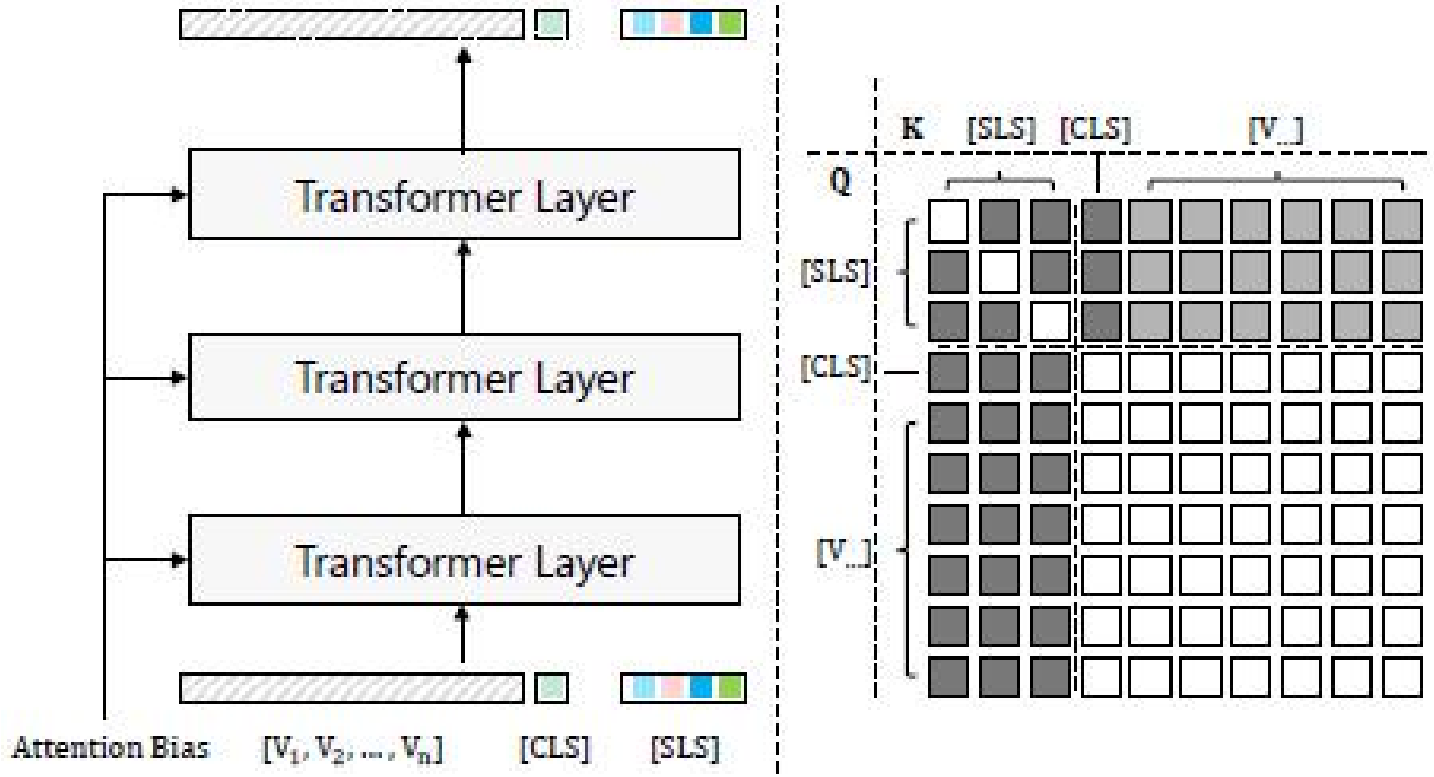


Xu, Mengde, et al. "Side adapter network for open-vocabulary semantic segmentation." CVPR 2023.

研究现状-知识获取-模型适配-SAN

方法-基于注意力倾向识别分割图：

- 不改变 CLIP 模型参数，引导[CLS]特征的注意力映射到正确区域
- 创建[CLS]特征的副本[SLS]特征
- [SLS]特征在注意力倾向的影响下**单向更新**，逐渐接近掩码预测
- 比较[SLS]特征与CLIP的文本特征之间的相似度得到掩码类别预测



Xu, Mengde, et al. "Side adapter network for open-vocabulary semantic segmentation." CVPR 2023.

研究现状-知识获取-模型适配-SAN

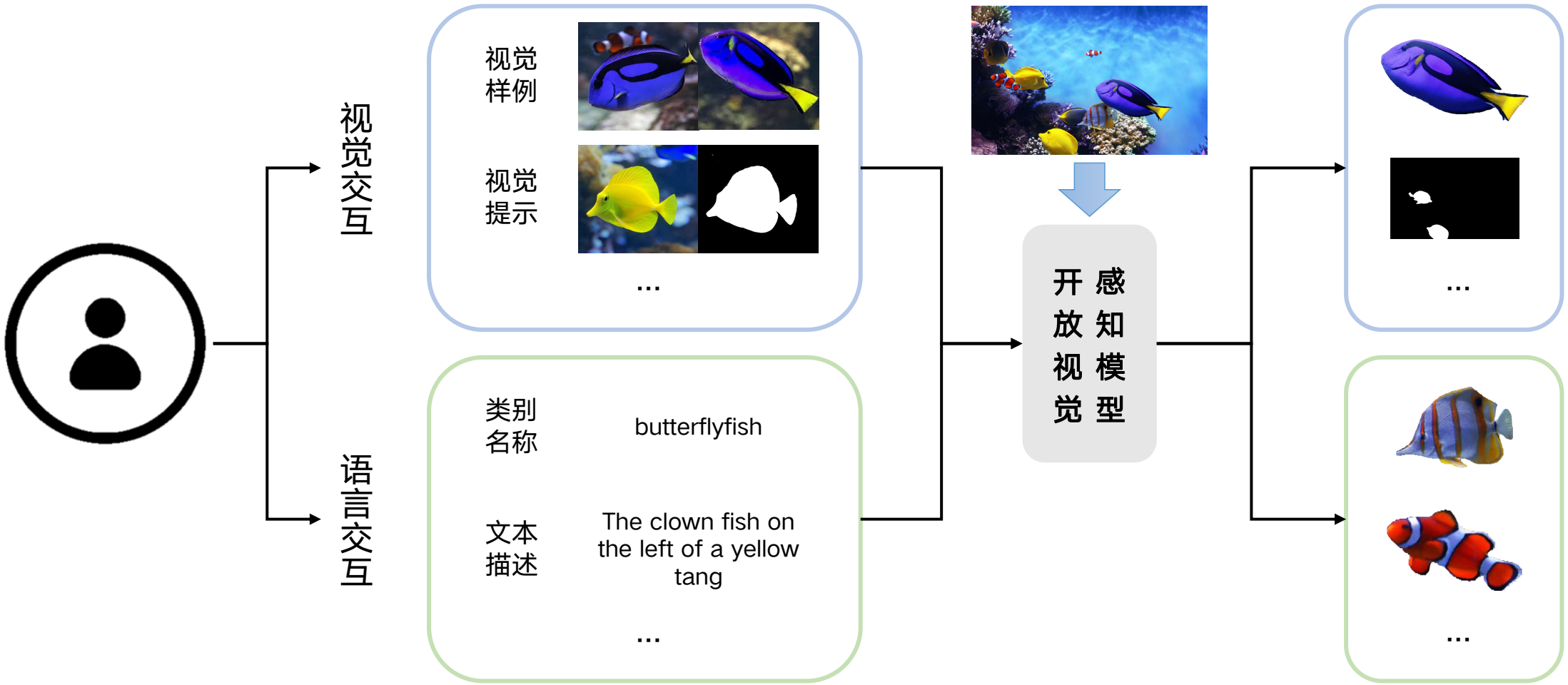
实验结果：

- 在ADE20K-150、Pascal Context-59等5个语义分割基准测试中取得**最先进性能**
- 具有**更少**的可训练参数和**更低**的计算成本

Xu, Mengde, et al. "Side adapter network for open-vocabulary semantic segmentation." CVPR 2023.

研究现状-概念理解

交互引导：运用知识，理解用户输入的概念，执行视觉感知

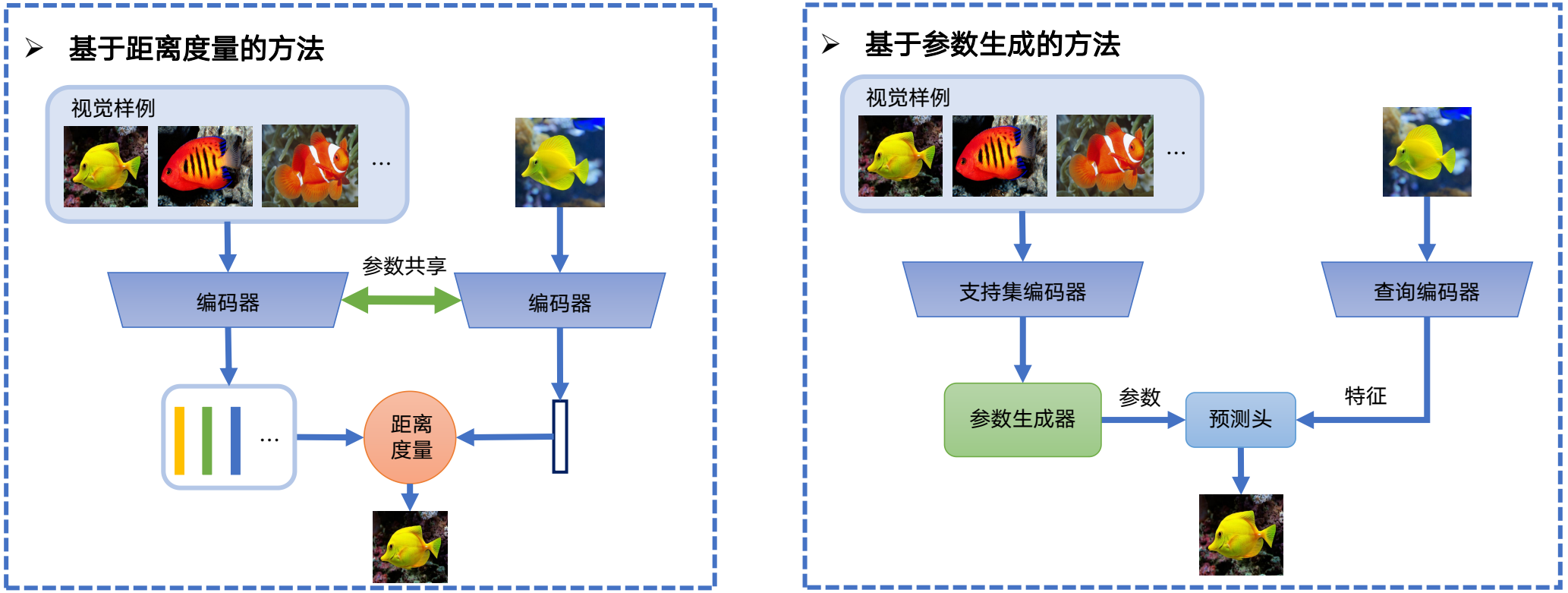


研究现状-概念理解-视觉样例

① 视觉样例

流程：利用交互给出的视觉样例与查询图片进行匹配，理解训练时未见过的概念

特点：概念描述直观具体，无需在视觉样例上进行微调



Yaqing Wang, et al. “Generalizing from a few examples: A survey on few-shot learning.” ACM computing surveys 2020.

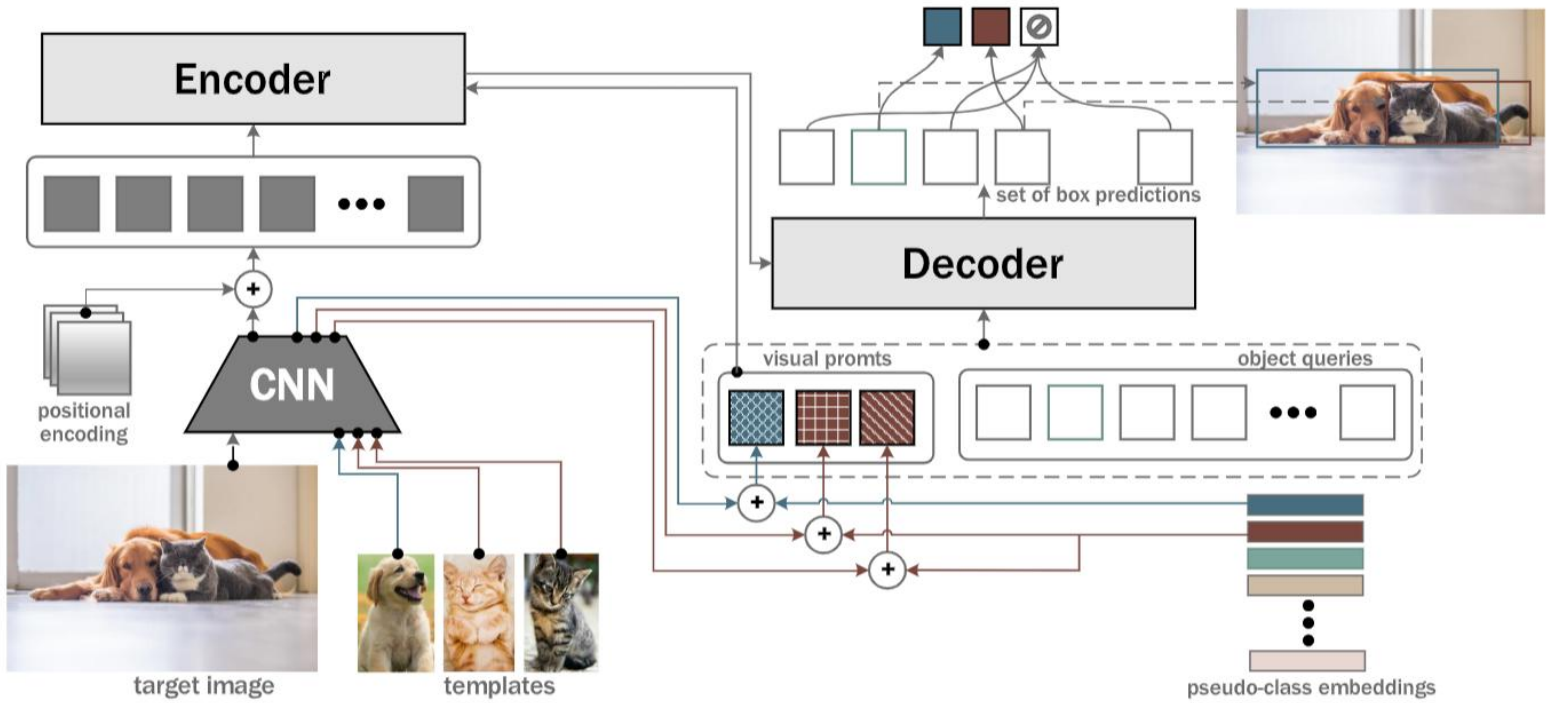
Wenqi Ren, et al. “Visual semantic segmentation based on few/zero-shot learning: An overview.” IEEE/CAA Journal of Automatica Sinica 2023.

研究现状-概念理解-视觉样例-FS-DETR

支持一次检测多个新对象，每个类别由可变数量的视觉样例表示

特点：

- 无需在视觉样例上微调
- 伪类别特征辅助提示
- 简单高效的无监督预训练方法

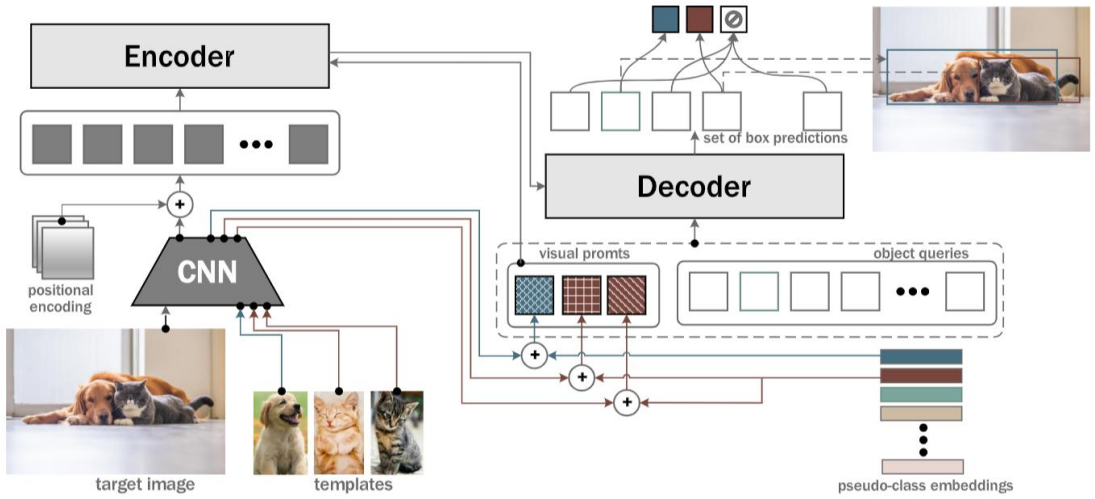


Adrian Bulat, et al. “FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training.” ICCV 2023.

研究现状-概念理解-视觉样例-FS-DETR

方法组成：

- CNN特征提取网络：用于从目标图像和视觉样例中提取视觉特征
- Transformer编码器：在图像特征内部进行自注意力，在图像特征和视觉样例之间进行交叉注意力机制
- Transformer解码器：输入目标索引和模板，预测伪类别和物体位置

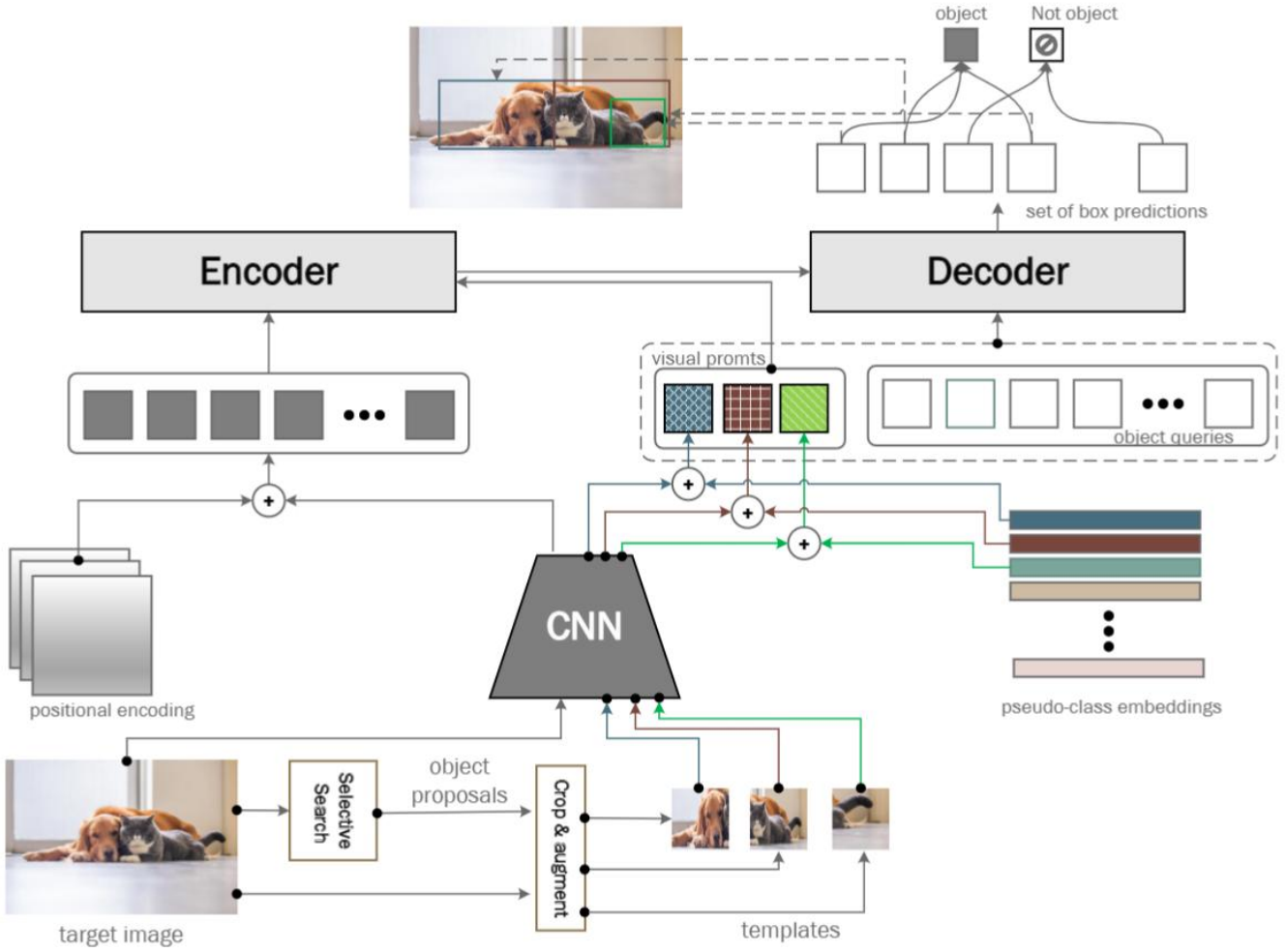


Adrian Bulat, et al.“FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training.” ICCV 2023.

研究现状-概念理解-视觉样例-FS-DETR

无监督预训练大致与训练阶段类似，主要有以下不同：

- 无需标注
- 目标边界框是通过选择性搜索或随机抽样提出的
- 模板从目标图像本身中采样
- 只定义前背景类



Adrian Bulat, et al.“FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training.” ICCV 2023.

研究现状-概念理解-视觉样例-FS-DETR

FS-DETR在少样本目标检测任务上具有显著优越性

Method / Shot	Venue	Backbone	Novel Set 1					Novel Set 2					Novel Set 3				
			1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Re-training based methods (meta-learning or fine-tuning)																	
FSRW* Kang et al. (2019)	ICCV'19	YOLOv2	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet* Wang et al. (2019)	ICCV'19	VGG16	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN* Yan et al. (2019)	ICCV'19	RN-101	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA w/cos Wang et al. (2020)	ICML'20	RN-101	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
TFA w/cos* Wang et al. (2020)	ICML'20	RN-101	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
Xiao *et al.* Xiao & Marlet (2020)	ECCV'20	RN-101	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
MPSR* Wu et al. (2020)	ECCV'20	RN-101	41.7	42.5	51.4	55.2	61.8	24.4	29.3	39.2	39.9	47.8	35.6	41.8	42.3	48.0	49.7
Fan *et al.*† Fan et al. (2020)	CVPR'20	RN-101	37.8	43.6	51.6	56.5	58.6	22.5	30.6	40.7	43.1	47.6	31.0	37.9	43.7	51.3	49.8
DCNET Hu et al. (2021)	CVPR'21	RN-101	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
TIP Li & Li (2021)	CVPR'21	RN-101	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9
CME Li et al. (2021)	CVPR'21	RN-101	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
SRR-FSD Zhu et al. (2021a)	CVPR'21	RN-101	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
Zhang *et al.** Zhang & Wang (2021)	CVPR'21	RN-101	47.0	44.9	46.5	54.7	54.7	26.3	31.8	37.4	37.4	41.2	40.4	42.1	43.3	51.4	49.6
QA-FewDet Han et al. (2021)	ICCV'21	RN-101	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5
DeFRCN Qiao et al. (2021)	ICCV'21	RN-101	40.2	53.6	58.2	63.6	66.5	29.5	39.7	43.4	48.1	52.8	35.0	38.3	52.9	57.7	60.8
DeFRCN* Qiao et al. (2021)	ICCV'21	RN-101	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4
Without re-training methods																	
Fan *et al.*† Fan et al. (2020)	CVPR'20	RN-101	32.4	22.1	23.1	31.7	35.7	14.8	18.1	24.4	18.6	19.5	25.8	20.9	23.9	27.8	29.0
UP-DETR‡ Dai et al. (2021)	ICCV'21	DETR-R50	38.2	40.4	44.5	45.8	46.0	20.0	23.6	25.8	28.0	33.9	34.1	35.3	37.0	40.1	40.3
QA-FewDet Han et al. (2021)	ICCV'21	RN-101	41.0	33.2	35.3	47.5	52.0	23.5	29.4	37.9	35.9	37.1	33.2	29.4	37.6	39.8	41.5
FS-DETR (Ours)	this work	DETR-R50	**45.0**	**48.5**	**51.5**	**52.7**	**56.1**	**37.3**	**41.3**	**43.4**	**46.6**	**49.0**	**43.8**	**47.1**	**50.6**	**52.1**	**56.9**

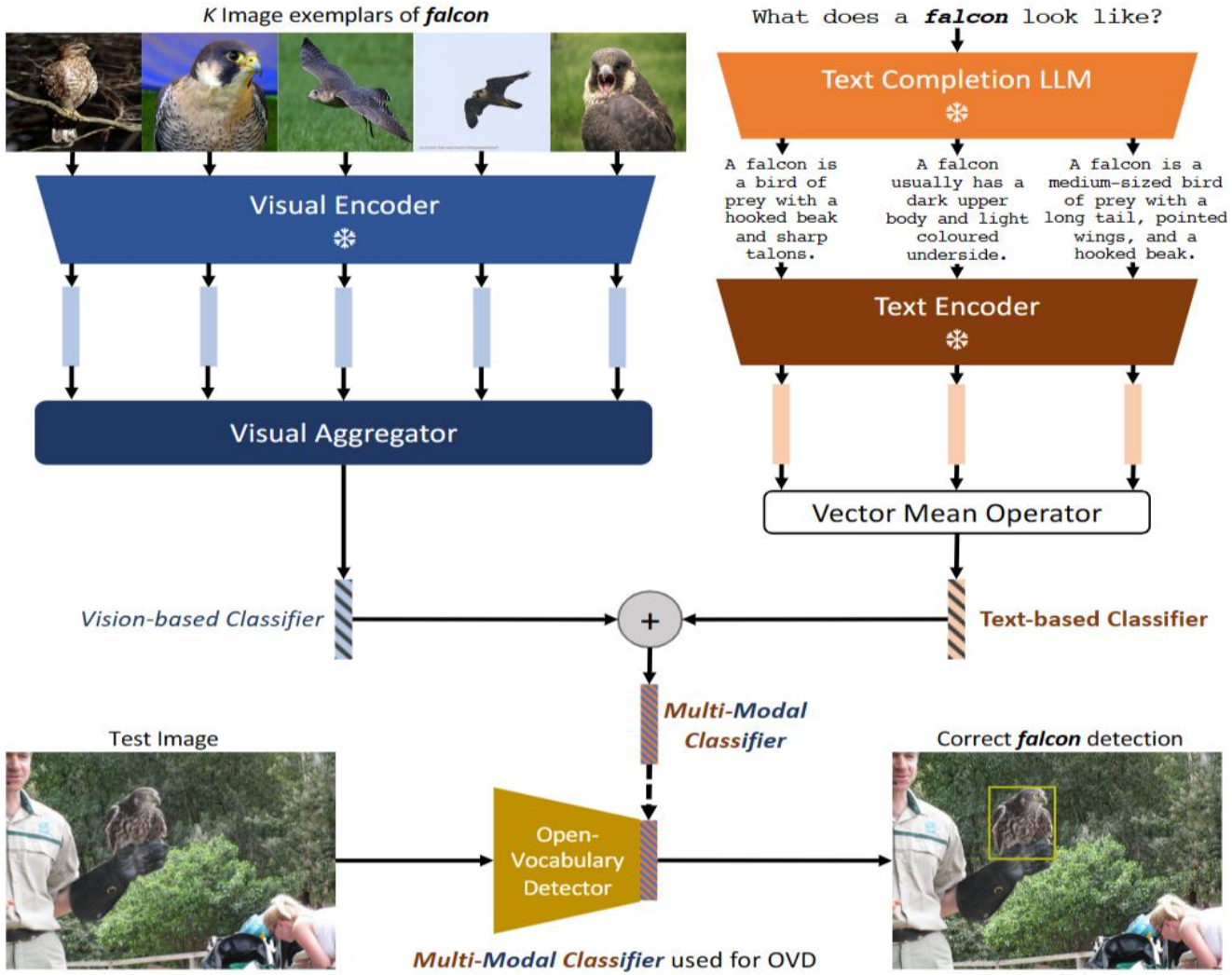
Method	1-shot			2-shot			3-shot			5-shot			10-shot		
	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Re-trained methods (meta-learning or fine-tuning)															
FSRW* Kang et al. (2019)	–	–	–	–	–	–	–	–	–	–	–	–	5.6	12.3	4.6
MetaDet* Wang et al. (2019)	–	–	–	–	–	–	–	–	–	–	–	–	7.1	14.6	6.1
Meta R-CNN* Yan et al. (2019)	–	–	–	–	–	–	–	–	–	–	–	–	8.7	19.1	6.6
TFA w/cos Wang et al. (2020)	1.9	3.8	1.7	3.9	7.8	3.6	5.1	9.9	4.8	7.0	13.3	6.5	9.1	17.1	8.8
TFA w/cos*† Wang et al. (2020)	3.4	5.8	3.8	4.6	8.3	4.8	6.6	12.1	6.5	8.3	15.3	8.0	10.0	19.1	9.3
Xiao *et al.*† Xiao & Marlet (2020)	3.2	8.9	1.4	4.9	13.3	2.3	6.7	18.6	2.9	8.1	20.1	4.4	10.7	25.6	6.5
MPSR*† Wu et al. (2020)	2.3	4.1	2.3	3.5	6.3	3.4	5.2	9.5	5.1	6.7	12.6	6.4	9.8	17.9	9.7
Fan *et al.*† Fan et al. (2020)	4.2	9.1	3.0	5.6	14.0	3.9	6.6	15.9	4.9	8.0	18.5	6.3	9.6	20.7	7.7
DCNET Hu et al. (2021)	-	-	-	-	-	-	-	-	-	-	-	-	12.8	23.4	11.2
TIP Li & Li (2021)	-	-	-	-	-	-	-	-	-	-	-	-	16.3	33.2	14.1
CME Hu et al. (2021)	-	-	-	-	-	-	-	-	-	-	-	-	15.1	24.6	16.4
SRR-FSD Zhu et al. (2021a)	-	-	-	-	-	-	-	-	-	-	-	-	11.3	23.0	9.8
Zhang *et al.* Zhang & Wang (2021)	4.4	7.5	4.9	5.6	9.9	5.9	7.2	13.3	7.4	-	-	-	-	-	-
QA-FewDet Han et al. (2021)	4.9	10.3	4.4	7.6	16.1	6.2	8.4	18.0	7.3	9.7	20.3	8.6	11.6	23.9	9.8
DeFRCN Qiao et al. (2021)	4.8	-	-	8.5	-	-	10.7	-	-	13.6	-	-	16.8	-	-
DeFRCN* Qiao et al. (2021)	9.3	-	-	12.9	-	-	14.8	-	-	16.1	-	-	18.5	-	-
Methods without re-training															
Fan *et al.*† Fan et al. (2020)	4.0	8.5	3.5	5.4	11.6	4.6	5.9	12.5	5.0	6.9	14.3	6.0	7.6	15.4	6.8
QA-FewDet Han et al. (2021)	5.1	10.5	4.5	7.8	16.4	6.6	8.6	17.7	7.5	9.5	19.3	8.5	10.2	20.4	9.0
FS-DETR (Ours)	**7.0**	**13.6**	**7.5**	**8.9**	**17.5**	**9.0**	**9.8**	**18.5**	**9.8**	**10.7**	**20.5**	**10.8**	**11.1**	**21.6**	**11.0**

Adrian Bulat, et al. "FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training." ICCV 2023.

研究现状-概念理解-视觉样例-MM-OVOD

探索定义新类别的三种方式：通过**语言描述**、通过**图像示例**或通过**两者组合**。

融合来自语言描述和图像样例的信息，产生多模态分类器。

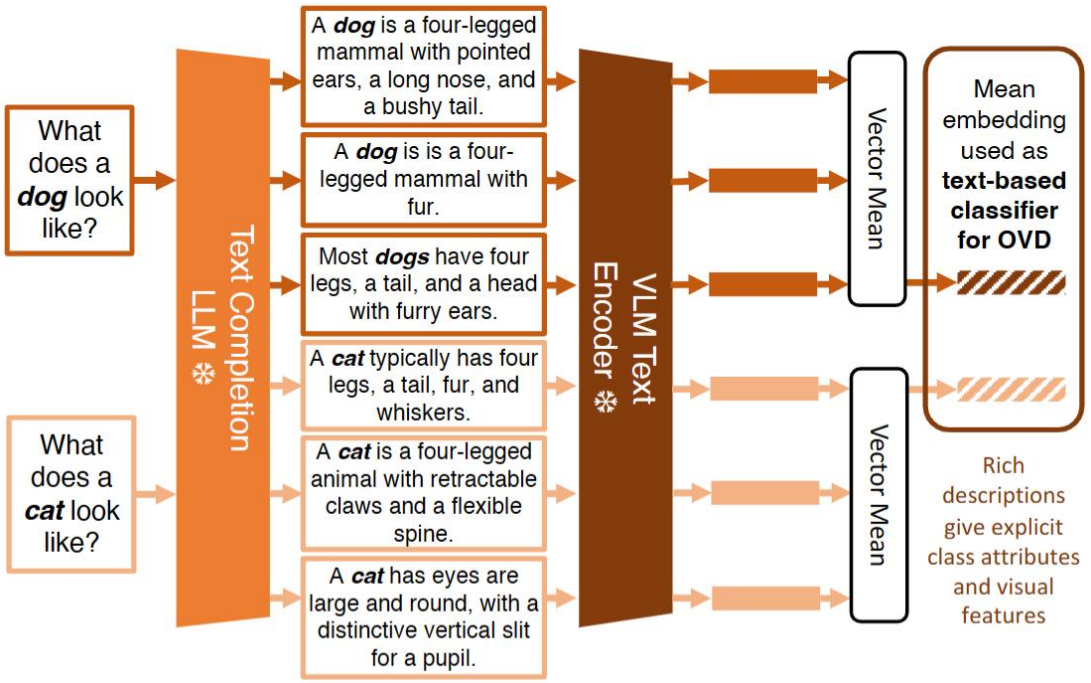


Prannay Kaul, et al. "Multi-Modal Classifiers for Open-Vocabulary Object Detection." ICML 2023.

研究现状-概念理解-视觉样例-MM-OVOD

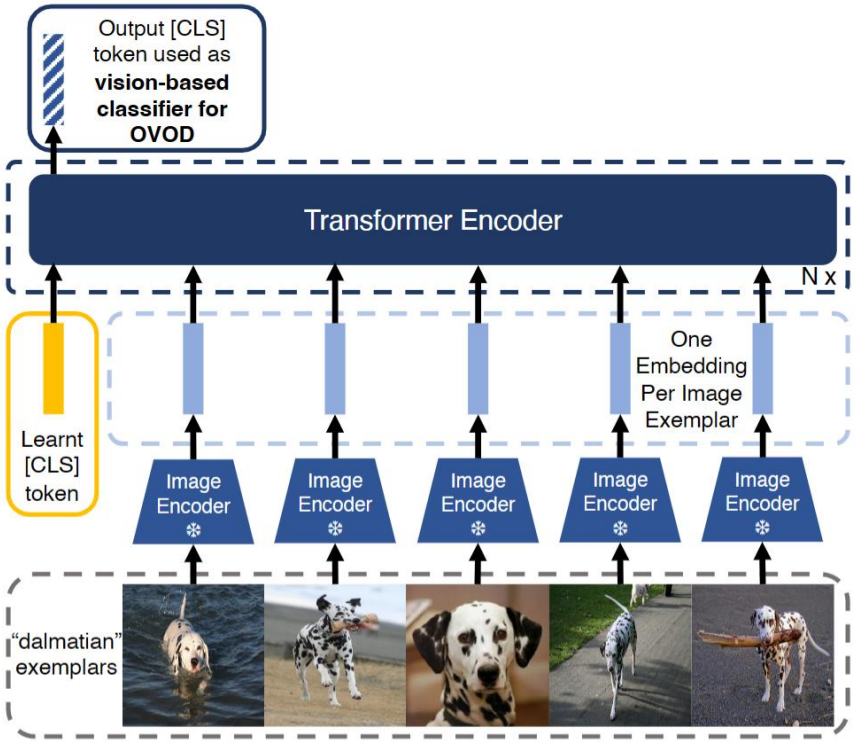
利用语言描述基于文本的分类器

- 提示大型语言模型为对象类生成信息丰富的语言描述



利用图像样例基于视觉的分类器:

- 在图像样本上使用视觉聚合器，摄取任意数量的图像作为输入，形成基于视觉的分类器



Prannay Kaul, et al. "Multi-Modal Classifiers for Open-Vocabulary Object Detection." ICML 2023.

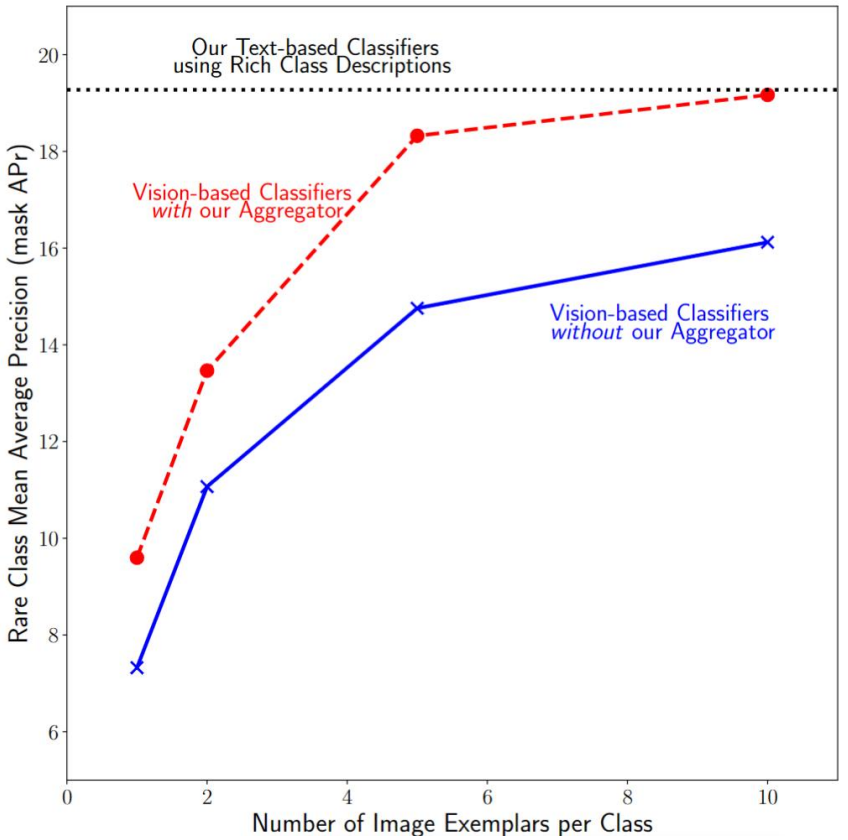
研究现状-概念理解-视觉样例-MM-OVOD

三种类型的分类器在LVIS开放词汇表检测基准上的检测性能

Model	Backbone	Extra Data	APr	mAP
ViLD (Gu et al., 2022)	ResNet-50		16.1	22.5
Detic (Zhou et al., 2022)	ResNet-50		16.3	30.0
ViLD-ens (Gu et al., 2022)	ResNet-50	✗	16.6	25.5
OV-DETR (Zang et al., 2022)	ResNet-50 + DETR		17.4	26.6
F-VLM (Kuo et al., 2022)	ResNet-50		*18.6*	24.2
Ours (Text-Based)			**19.3**	*30.3*
Ours (Vision-Based)	ResNet-50	✗	18.3	29.2
Ours (Multi-Modal)			**19.3**	**30.6**
RegCLIP (Zhong et al., 2022)	ResNet-50	CC3M	17.1	28.2
OWL-ViT (Minderer et al., 2022)†	ViT-B/32	LiT	19.7	23.3
Detic (Zhou et al., 2022)	ResNet-50	IN-L	24.6	32.4
Ours (Text-Based)			*25.8*	*32.7*
Ours (Vision-Based)	ResNet-50	IN-L	23.8	31.3
Ours (Multi-Modal)			**27.3**	**33.1**
Fully-Supervised (Zhou et al., 2022)	ResNet-50	✗	25.5	31.1

视觉样例数对视觉分类器性能的影响：

- 改变每个类可用的图像样本数量，以研究图像样本数量对OVOD性能的影响



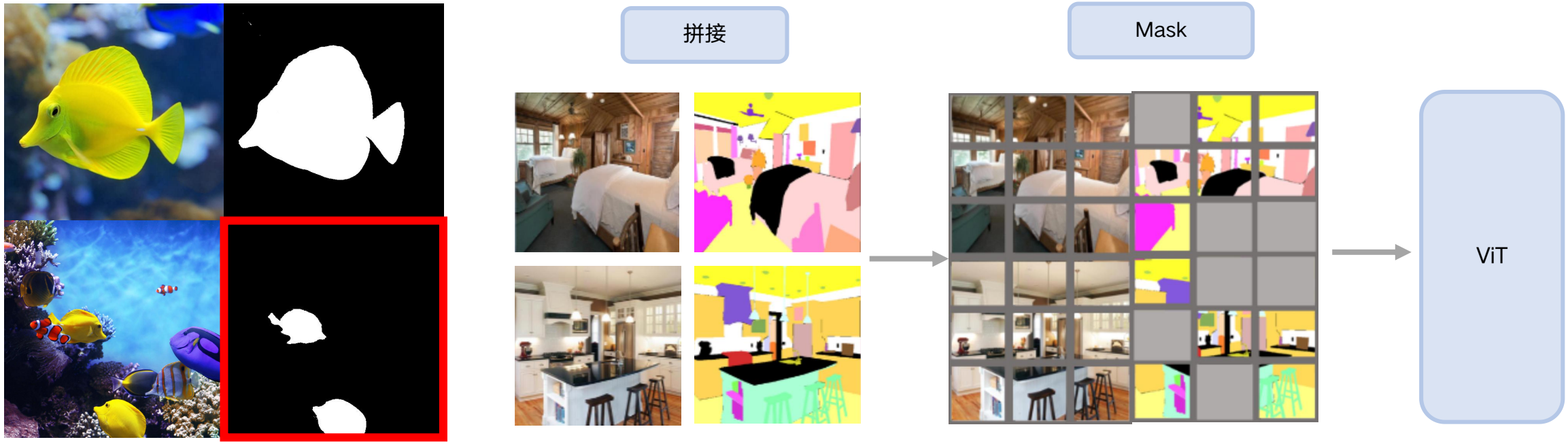
Prannay Kaul, et al. “Multi-Modal Classifiers for Open-Vocabulary Object Detection.” ICML 2023.

研究现状-概念理解

② 视觉提示

流程：提供输入输出样例作为提示，使模型快速适应各种任务

特点：无需接入特定任务预测头微调



Xinlong Wang, et al. "Images speak in images: A generalist painter for in-context visual learning." CVPR 2023.

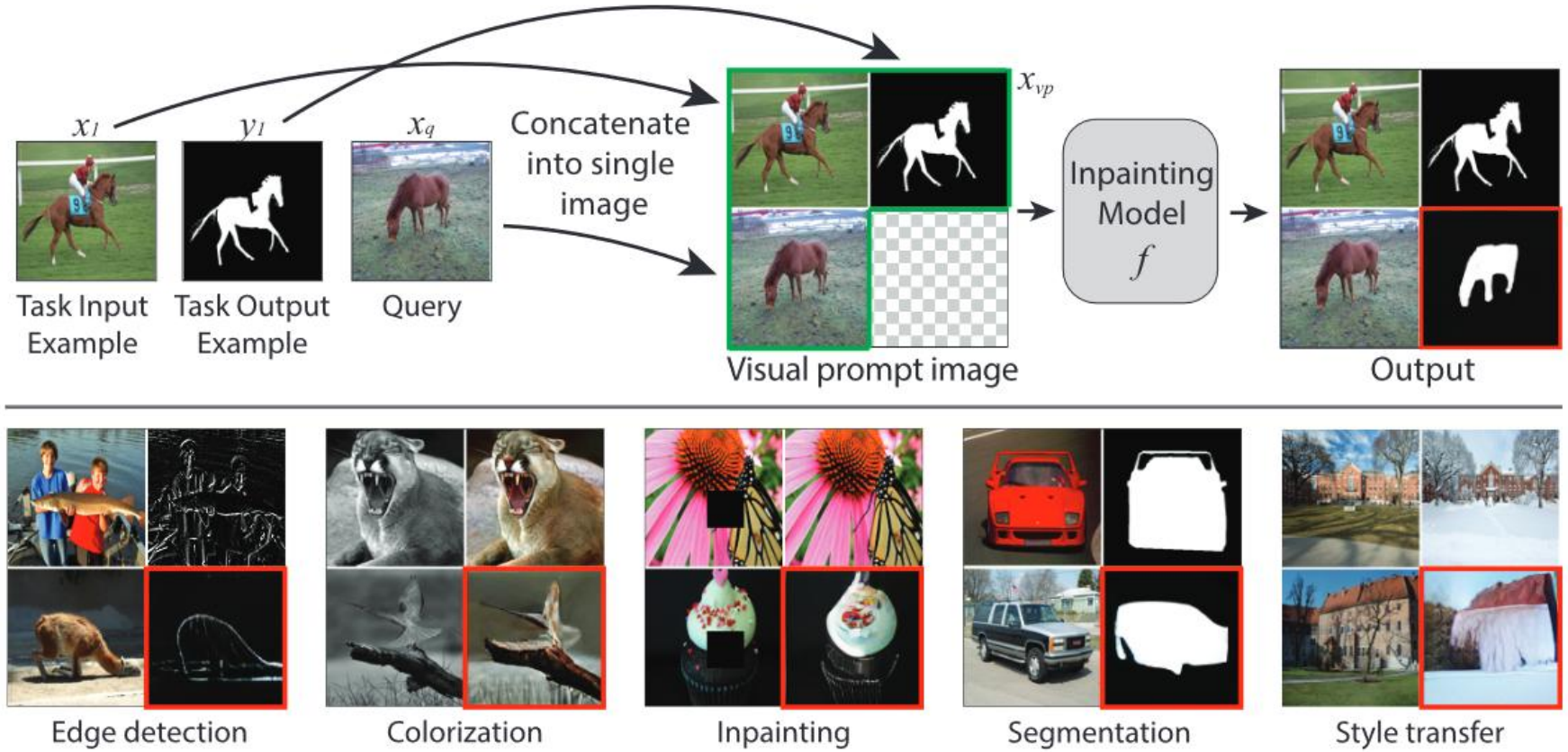
Xinlong Wang, et al. "Seggpt: Segmenting everything in context.”, ICCV 2023

研究现状-概念理解-视觉提示-VP

根据图片上下文学习能力，用 few-shot 迁移到下游任务

特点：

- 将多个任务统一为一个任务方式
- 所有任务都是“在多个连接图像里的空白处涂色”

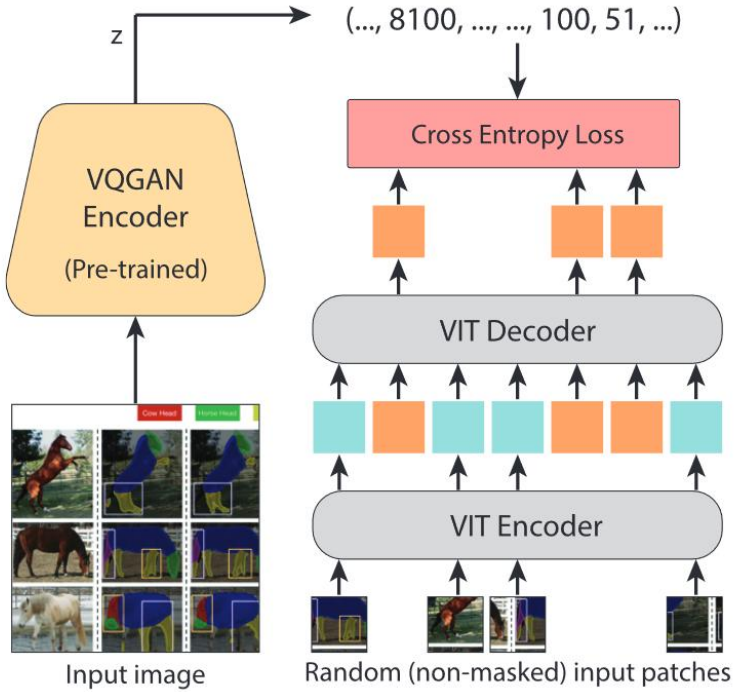


Amir Bar, et al. "Visual prompting via image inpainting." NeurIPS 2022.

研究现状-概念理解-视觉提示-VP

训练过程

- 采用 **MAE-VQGAN** 进行**掩码图像建模**
- 为了解决拼接图像与自然图像域不同，从 **arXiv** 的 **CV** 领域论文收集**拼接图像**训练

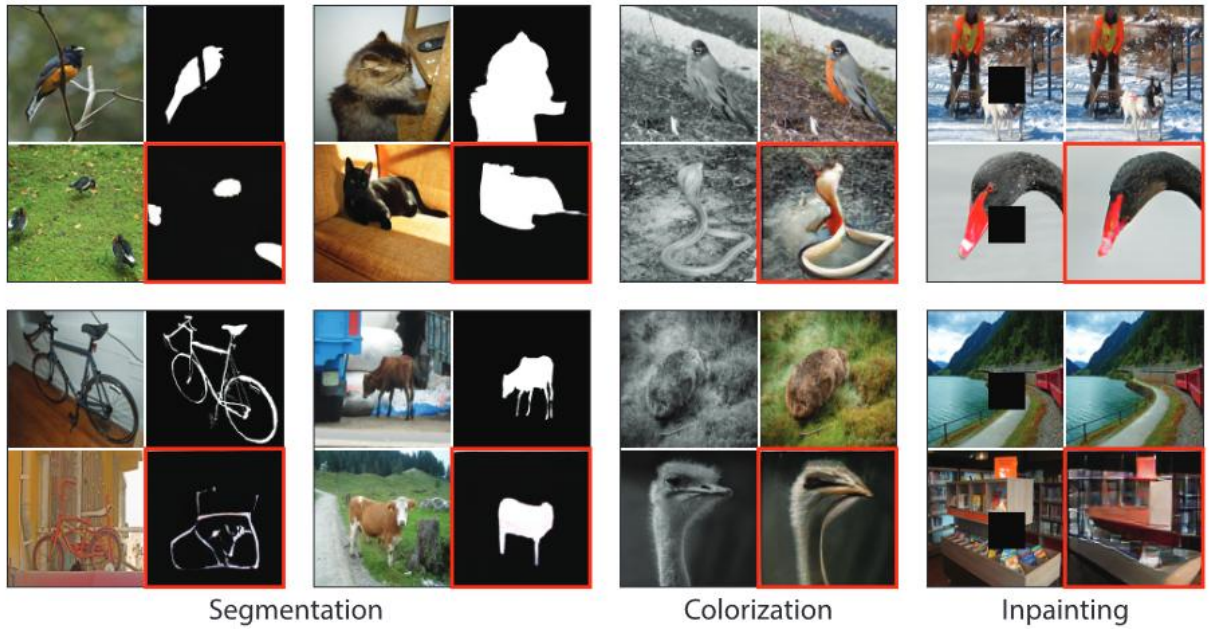


Amir Bar, et al. "Visual prompting via image inpainting." NeurIPS 2022.

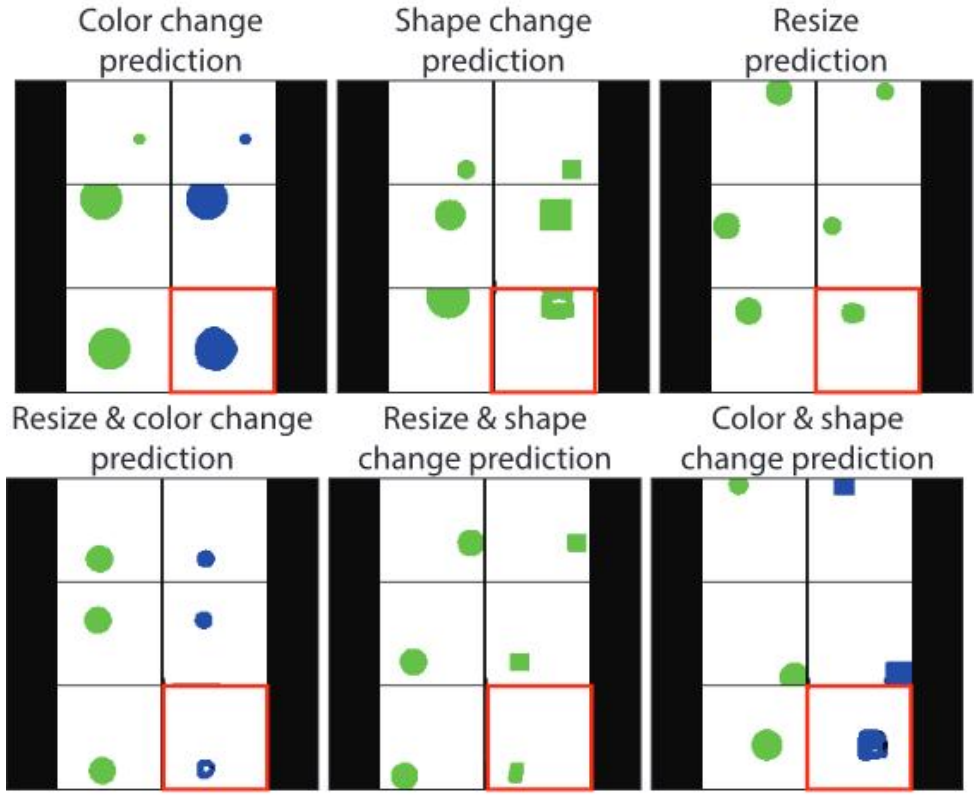
研究现状-概念理解-视觉提示-VP

实验结果

➢ 在多个不同下游任务的效果：



➢ 合成图像上的效果：



Amir Bar, et al. "Visual prompting via image inpainting." NeurIPS 2022.

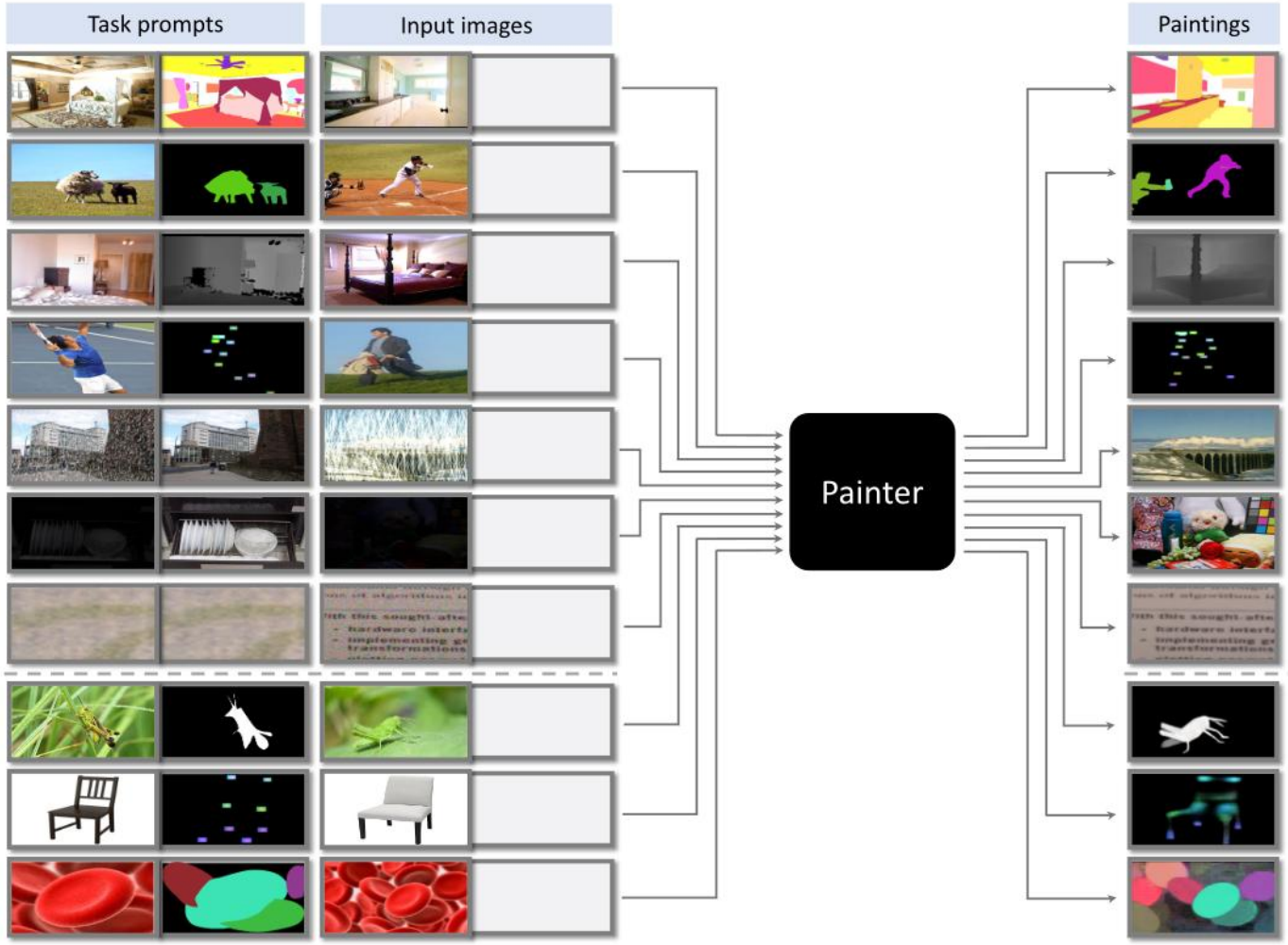
研究现状-概念理解-视觉提示-Painter

模型方案

- 发掘上下文学习能力 + MIM 训练
- 训练不同的 prompt 用于不同任务
- 需要设计统一的输出形式

输出形式统一化

- 统一输出空间到 RGB 图像
- 如深度估计将 0-10映射到 0-255 之间向下取整，语义分割的类别映射到具体的 RGB 值。

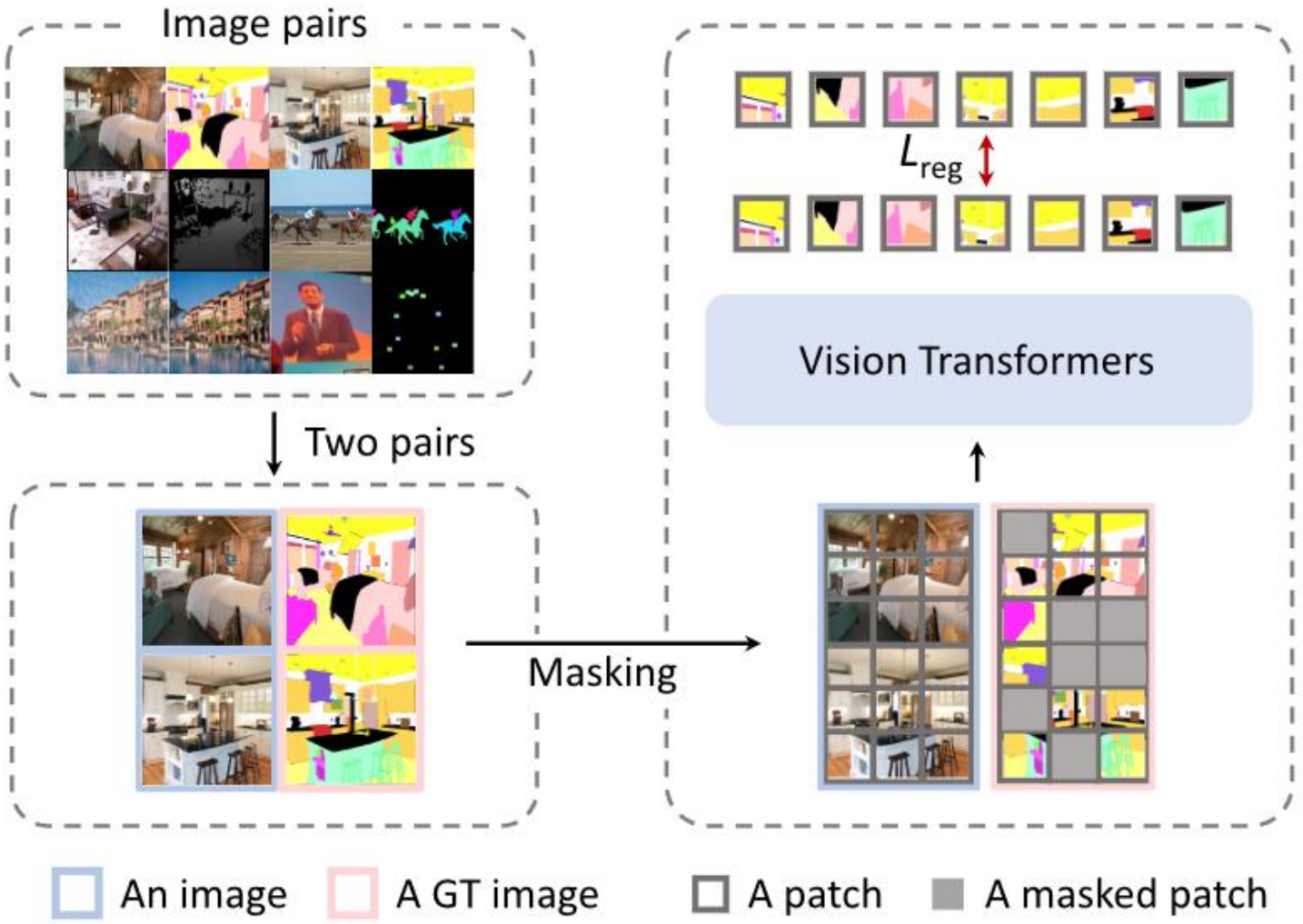


Xinlong Wang, et al. "Images speak in images: A generalist painter for in-context visual learning." CVPR 2023.

研究现状-概念理解-视觉提示-Painter

MIM 训练过程

- 从数据集中选择同域同任务的两个图像对，将图像对拼接在一起
- 随机对输出 mask 掉一些 patch
- 将 mask 的 patch 替换为可学习的向量，并将 mask token 经过 ViT 预测原 patch
- 用 smooth-L1 loss 计算损失



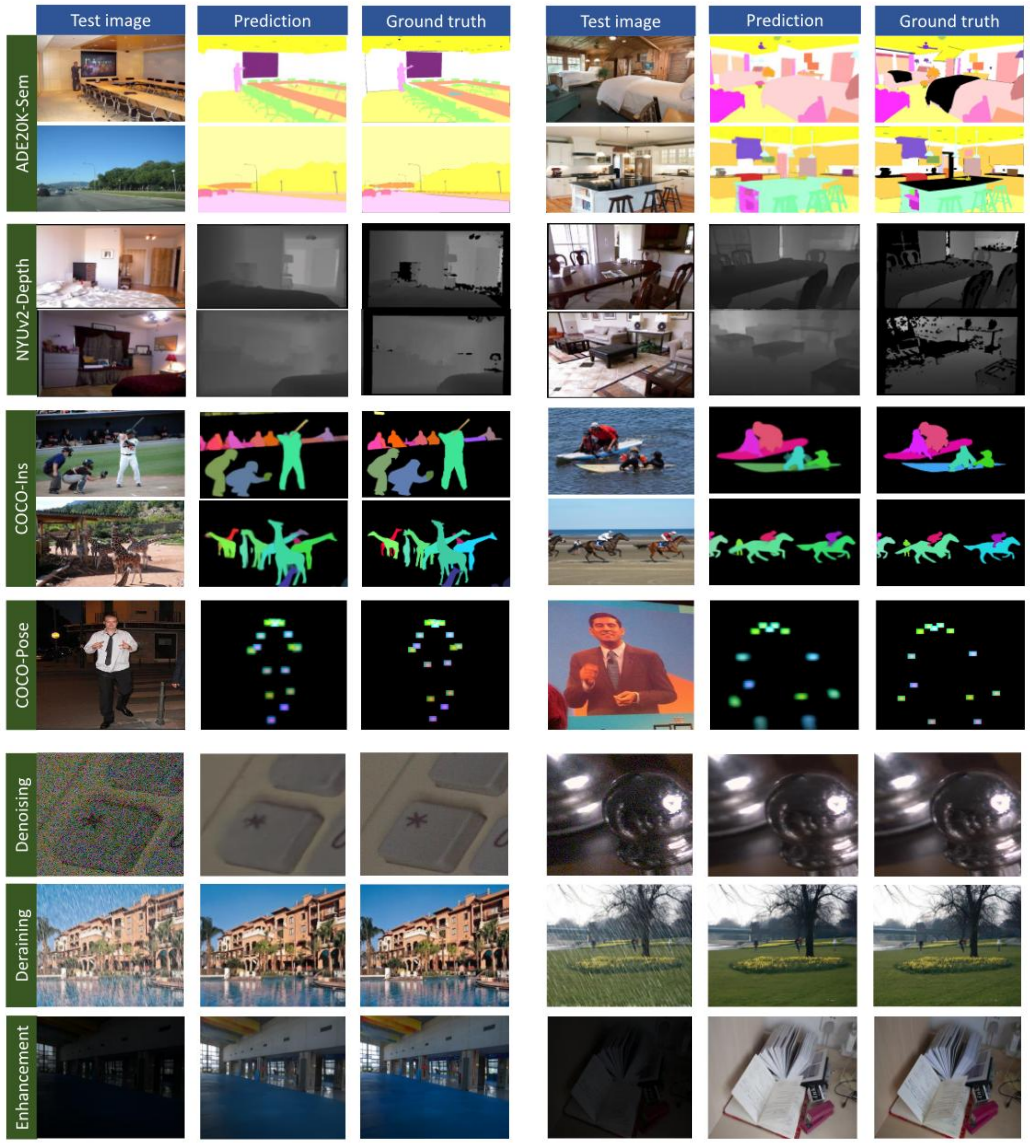
Xinlong Wang, et al. "Images speak in images: A generalist painter for in-context visual learning." CVPR 2023.

研究现状-概念理解-视觉提示-Painter

实验结果

在多个下游任务上都获得了不错的效果

	depth estimation NYUv2			semantic seg. ADE-20K	panoptic seg. COCO	keypoint det. COCO	denoising SIDD		deraining 5 datasets		enhance. LoL	
	RMSE ↓	A.Rel ↓	δ_1 ↑	mIoU ↑	PQ ↑	AP ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
specialized models												
DenseDepth [3]	0.465	0.123	0.846	-	-	-	-	-	-	-	-	-
BinsFormer [30]	0.330	0.094	0.925	-	-	-	-	-	-	-	-	-
UperNet-ViT-L [47]	-	-	-	49.9	-	-	-	-	-	-	-	-
Mask2Former [11]	-	-	-	57.7	57.8	-	-	-	-	-	-	-
DETR [8]	-	-	-	-	45.6	-	-	-	-	-	-	-
HRNet [46]	-	-	-	-	-	76.3	-	-	-	-	-	-
HRFormer [11]	-	-	-	-	-	77.2	-	-	-	-	-	-
Uformer [44]	-	-	-	-	-	-	39.89	0.960	-	-	-	-
MPRNet [50]	-	-	-	-	-	-	39.71	0.958	32.73	0.921	-	-
MIRNet-v2 [51]	-	-	-	-	-	-	39.84	0.959	-	-	24.74	0.851
generalist framework, specialized models												
UViM [27]	0.467	-	-	-	45.8	-	-	-	-	-	-	-
generalist models												
Unified-IO [35]	0.385	-	-	-	-	-	-	-	-	-	-	-
Pix2Seq v2 [10]	-	-	-	-	-	64.8	-	-	-	-	-	-
Painter (ours)	0.288	0.080	0.950	49.9	43.4	72.1	38.88	0.954	29.49	0.868	22.40	0.872

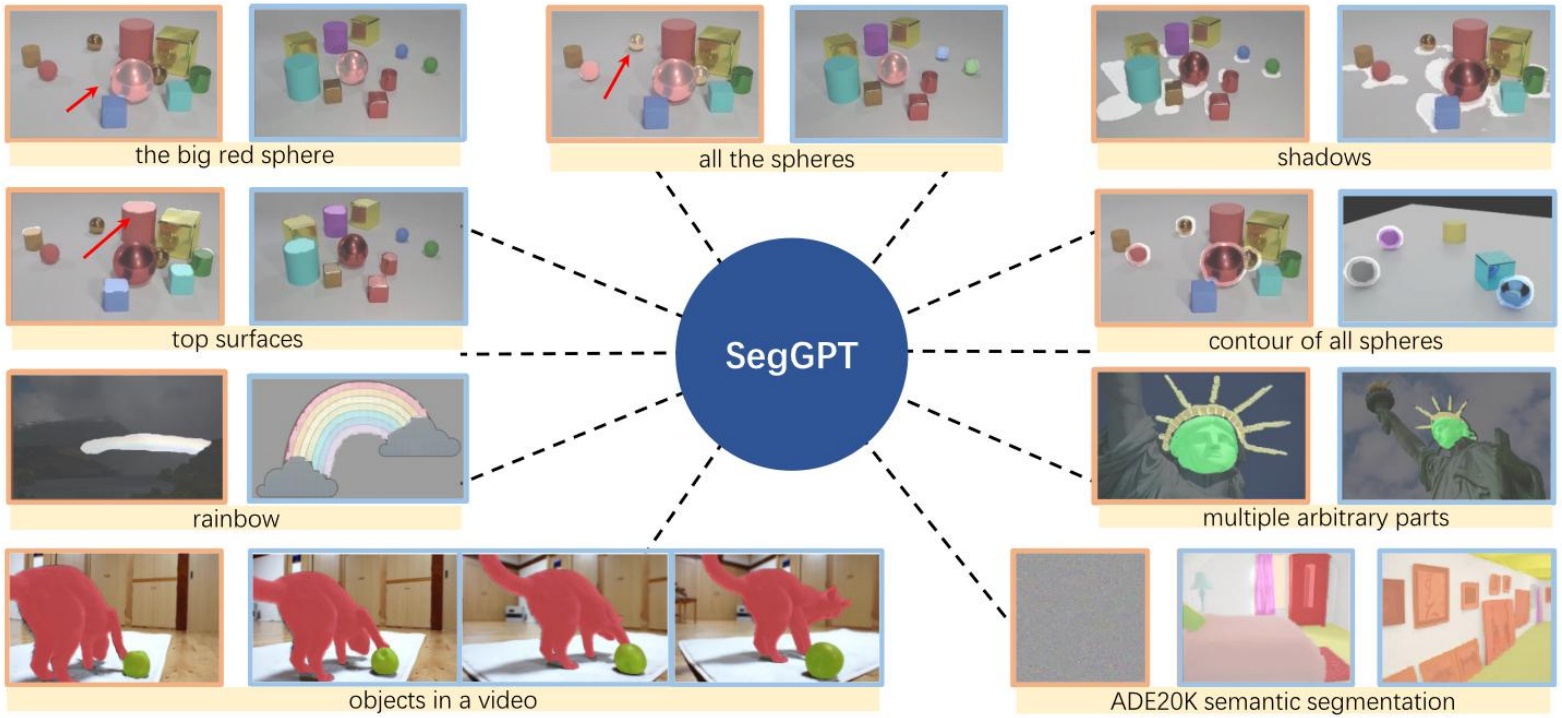


Xinlong Wang, et al. "Images speak in images: A generalist painter for in-context visual learning." CVPR 2023.

研究现状-概念理解-视觉提示-SegGPT

整体方案

- 模型基于 Painter，但是聚焦于分割任务
- Painter 的任务输出是预定义的，如语义分割按语义定义颜色，实例分割按位置定义颜色
- 为了避免模型可能会根据颜色方案知道具体任务，采用随机着色的方案



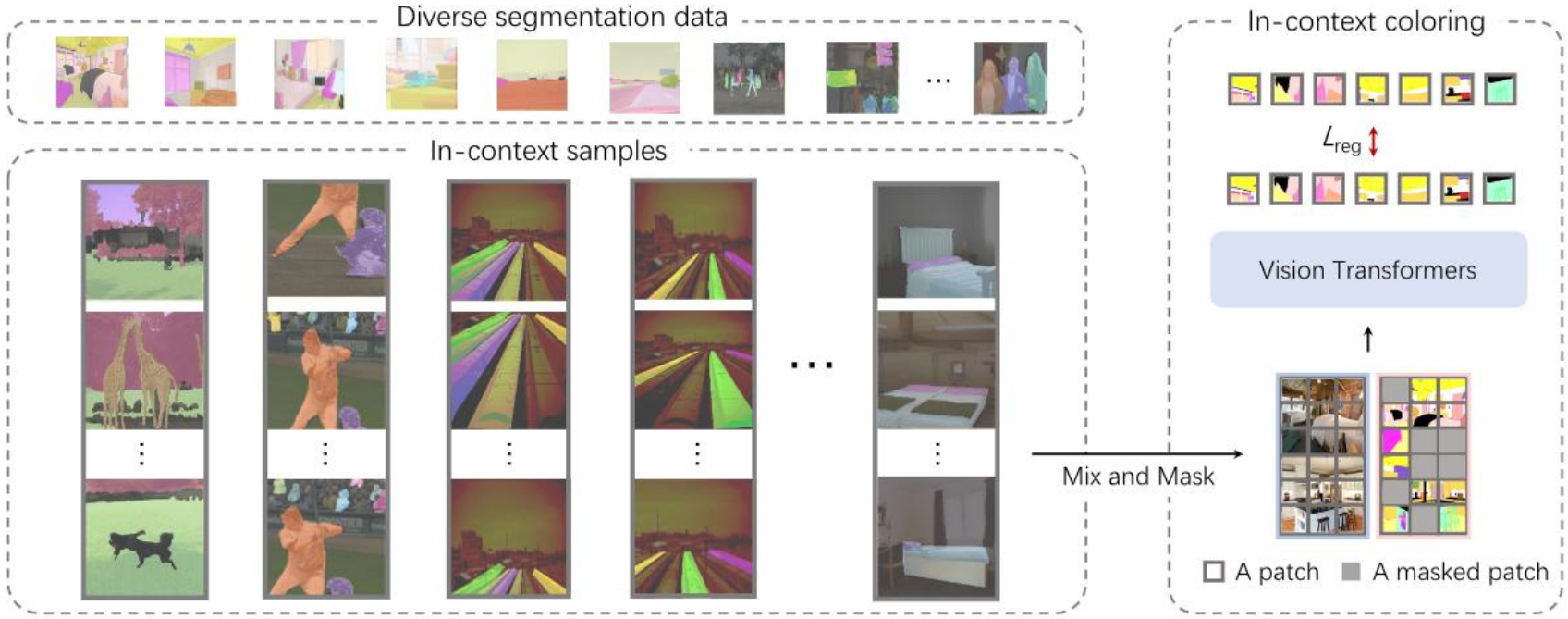
Xinlong Wang, et al. "Seggpt: Segmenting everything in context.”, ICCV 2023

研究现状-概念理解-视觉提示-SegGPT

MIM训练过程

- 将具有相同配色方案的多组数据组成in-context sample
- 除了从数据集中选择近似的图片，还使用随机裁剪与缩放的数据增广方案

- 对于一个图像，随机选另一幅具有相似的语义类别或者对象实例的图像，再随机采样一组颜色方案为两幅图像着色。

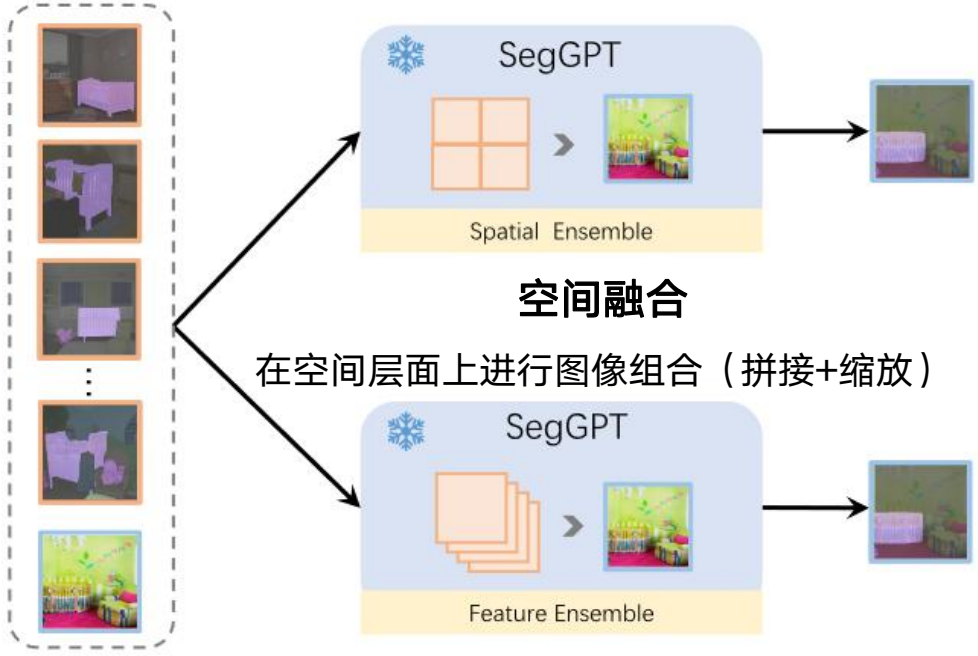


Xinlong Wang, et al. "Seggpt: Segmenting everything in context.”, ICCV 2023

研究现状-概念理解-视觉提示-SegGPT

多图方案设计

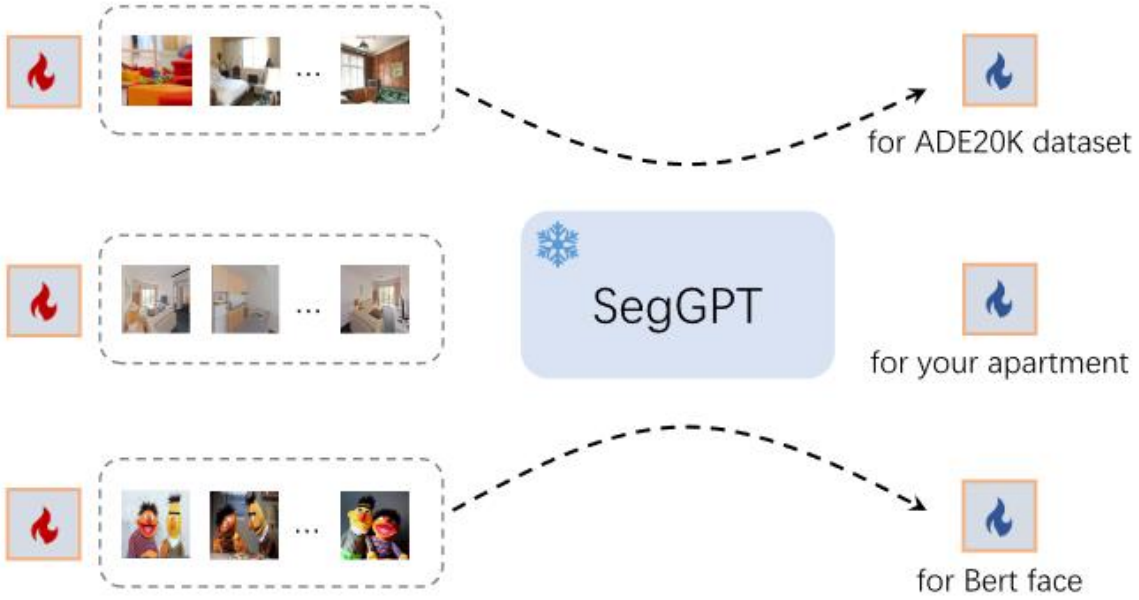
➢利用**多幅 prompt** 进一步提升分割准确性



经过每层 attn 时将 query image 特征平均一次

In-context Tuning

➢插入**可学习的图像向量**，在不同的下游任务上微调。

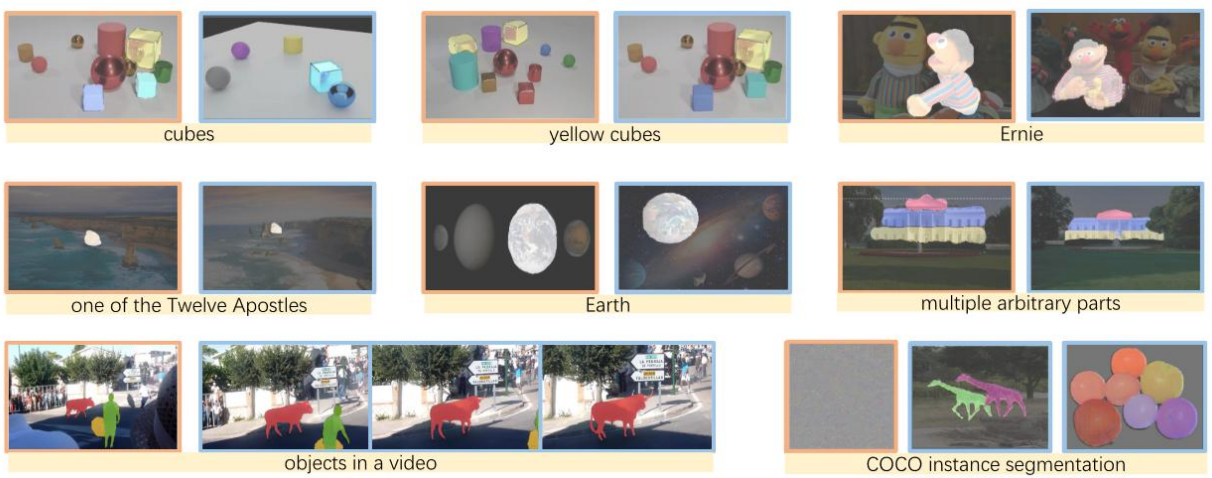


Xinlong Wang, et al. "Seggpt: Segmenting everything in context.”, ICCV 2023

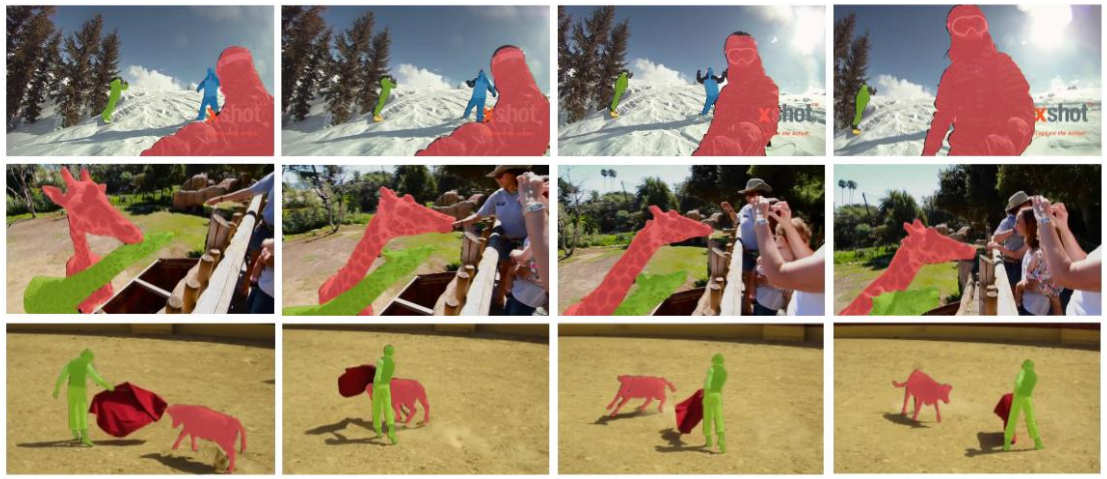
研究现状-概念理解-视觉提示-SegGPT

结果展示

- 在多个下游任务上都获得了不错的效果，如下图所示。



图像分割任务



视频分割任务

Xinlong Wang, et al. "Seggpt: Segmenting everything in context.”, ICCV 2023

研究现状-概念理解

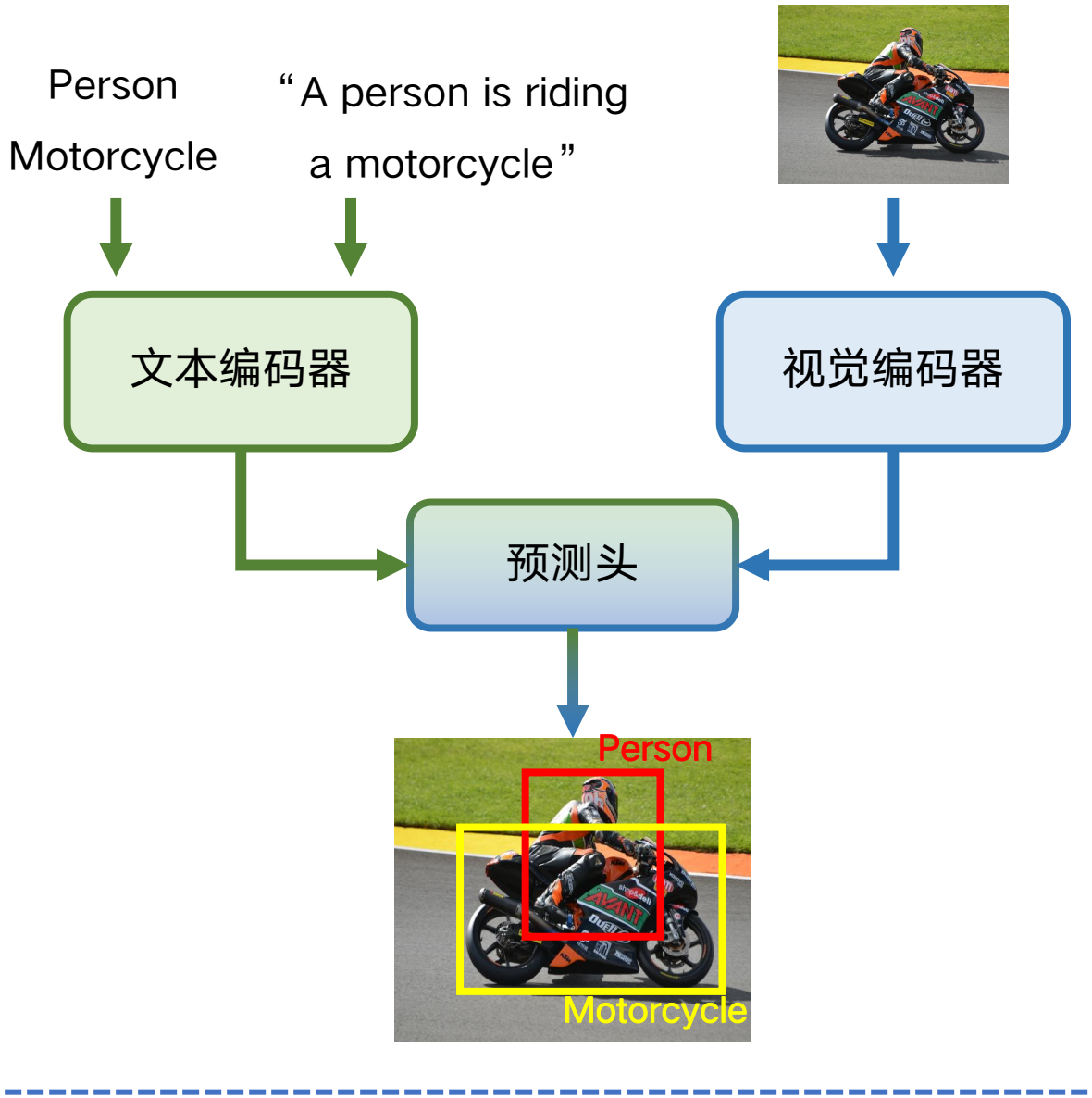
③ 自然语言

流程：对自然语言和图像分别编码后，在图像中**感知语言所指代的视觉内容**

特点：语言指令的**形式灵活多样**

类别标签 → 开放词表检测

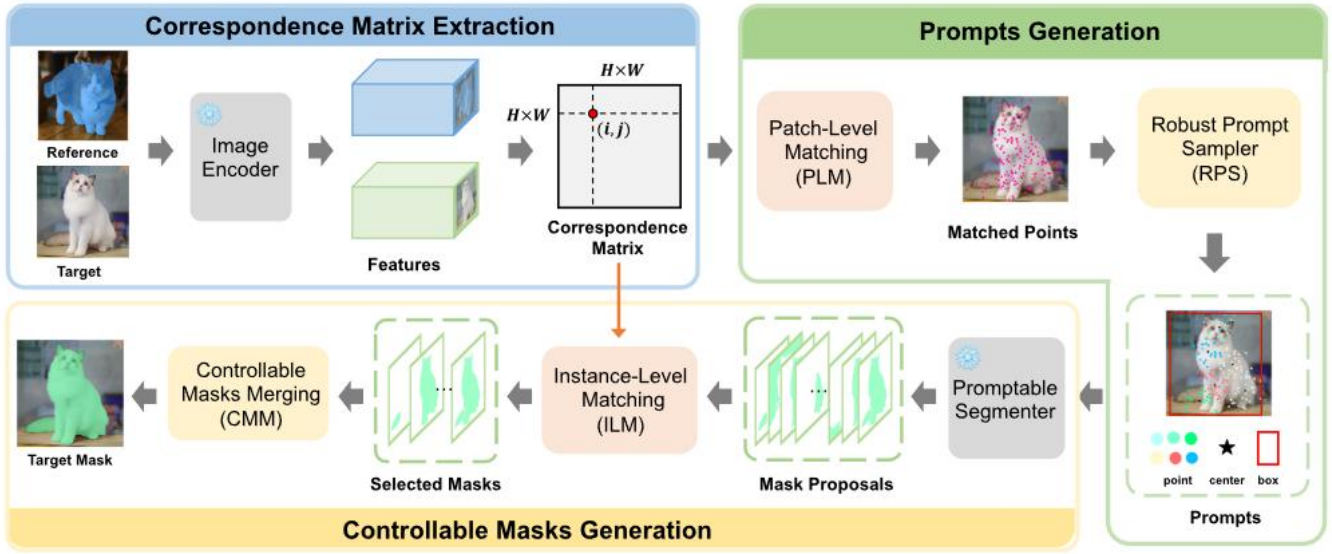
语句描述 → 短语目标定位
指代表达检测



研究现状-概念理解-自然语言-Matcher

Matcher: One-shot SAM

- 对于一个给定的**参考图像**及其 Mask，Matcher 可以在目标图像中分割**具有相同语义**信息的目标或者部位。
- 由于 SAM 的分割结果**缺乏语义信息**，且分割结果以**模棱两可**的 Mask 呈现，通过采用 prompt + SAM 辅助分割的方案解决该问题



Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

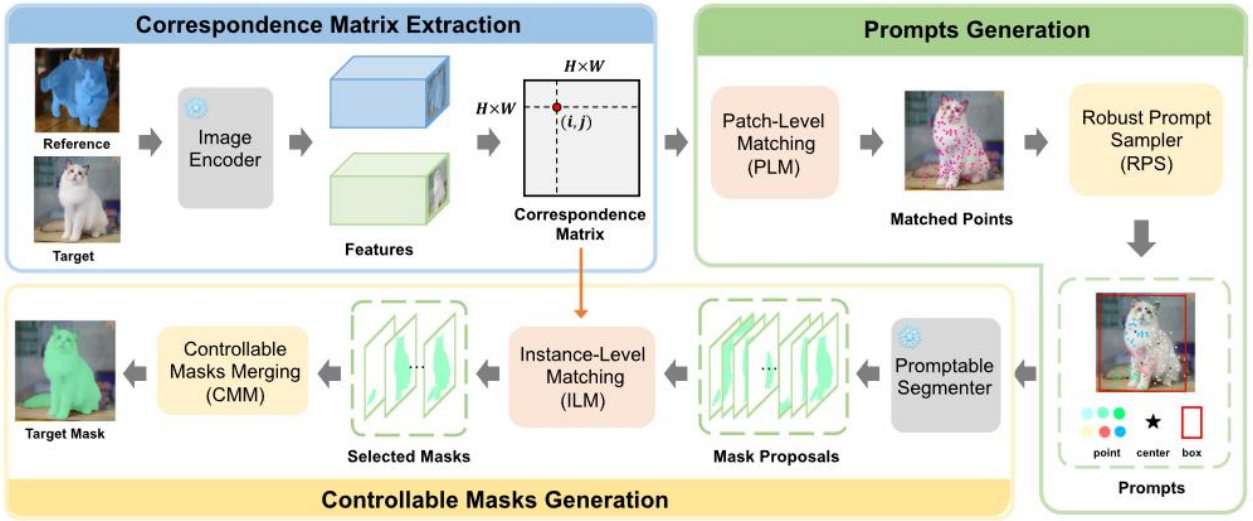
研究现状-概念理解-自然语言-Matcher

整体流程设计

- 通过计算图像特征之间的的**相似度**，来提取 Correspondence Matrix。
- 进行 Patch-Level Matching (PLM) **获取匹配点**，再采用Robust Prompt Sampler来**采样部分质量高的匹配点**，用于生成一些prompts（Point、Center以及Box）。
- 将上述 prompts 输入到SAM中，生成初始的Mask Proposals。
- 进行参考图像和Mask Proposals之间的Instance-Level Matching (ILM)，以**选取高质量的Masks**
- 使用Controllable Masks Merging (CMM) **合成最终的 Mask**。

主要部分

- Correspondence Matrix Extraction
- Prompts Generation
- Controllable Masks Generation

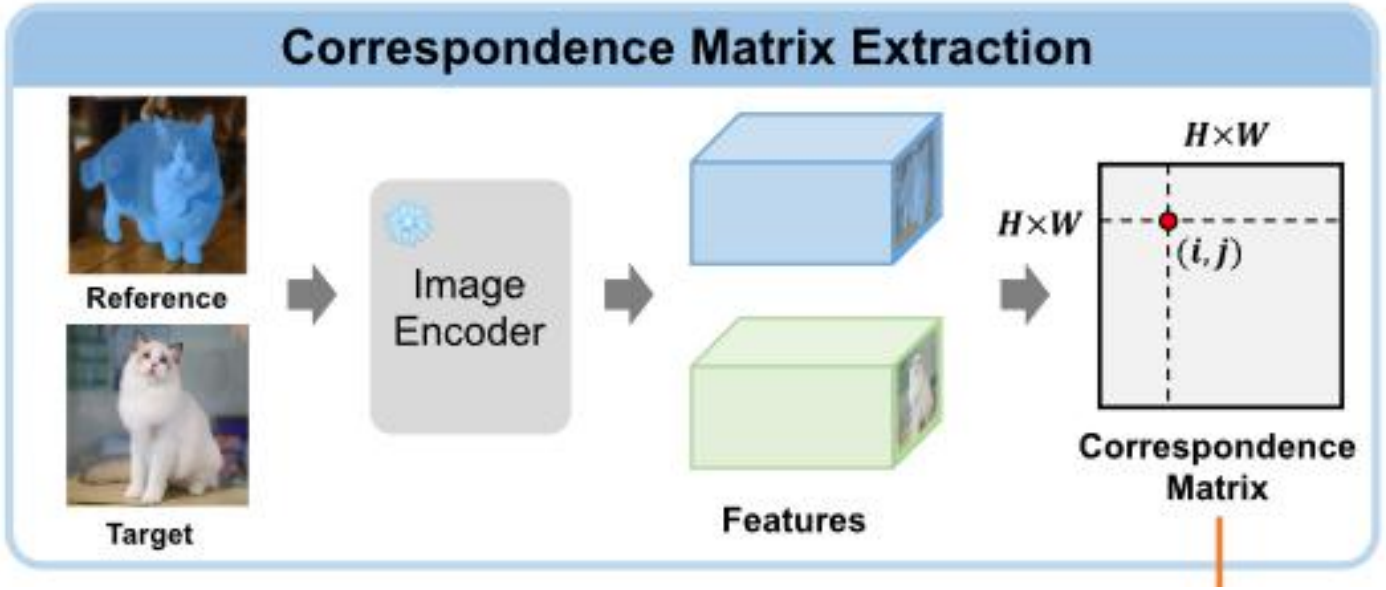


Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

研究现状-概念理解-自然语言-Matcher

Correspondence Matrix Extraction (CME)

- 计算两个特征 Patch-wise 的相似度来探索目标图像上参考 Mask 的**最佳匹配区域**。
- 对于图片特征 $z_r, z_t \in \mathbb{R}^{H \times W \times C}$，计算 Patch-wise 的**余弦相似度** $sim(z_r, z_t)$ 用于后续匹配。

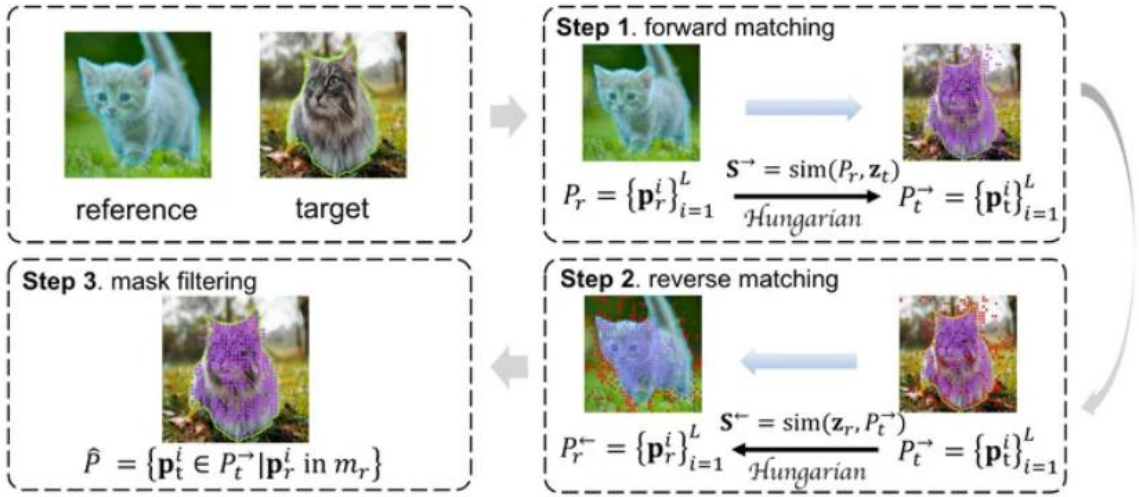


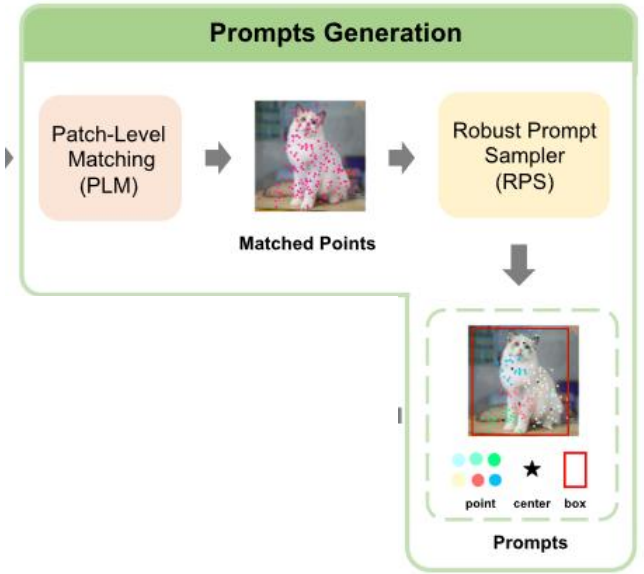
Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

研究现状-概念理解-自然语言-Matcher

Patch-Level Matching (PLM)

- 匹配 reference 与 target 的像素点。在一些困难场景，例如相似的上下文信息、多个实例等，Image Encoder 可能会预测一些错误的匹配结果。
- 只选择那些正向匹配 + 逆向匹配后还在 mask 里的点作为参考点。



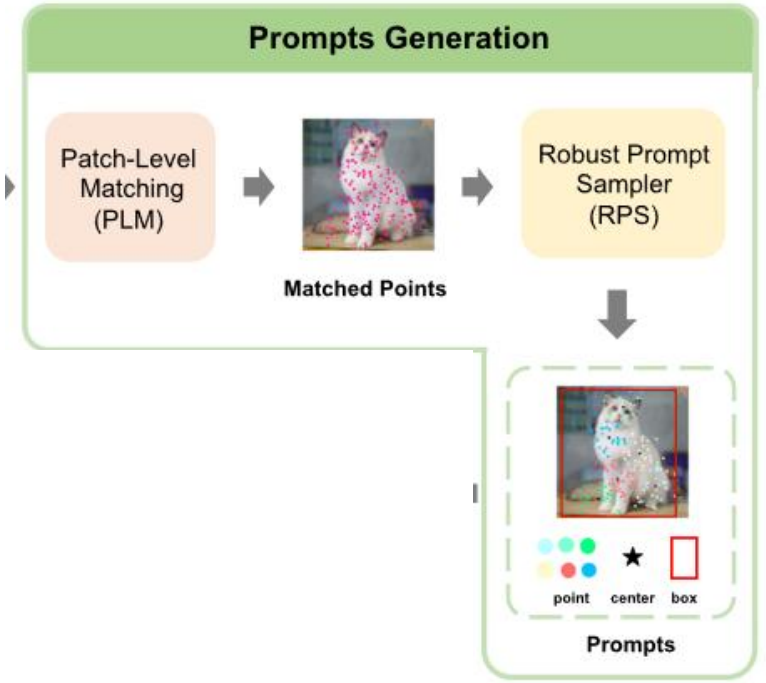


Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

研究现状-概念理解-自然语言-Matcher

Robust Prompt Sampler (RPS)

- 生成多样且有意义的 Mask Proposals，在不同语义粒度（部分、整体以及多实例）上实现鲁棒的分割，抑制由匹配异常值引起的碎片化假阳性 Mask 预测。



k-means++ 将匹配点分类为多个簇，并分为三类采样点：

- Patch-Level Prompts： 在每个簇内采样
- Instance-Level Prompts： 在所有匹配点内采样
- Global Prompts： 在所有簇中心点内进行采样

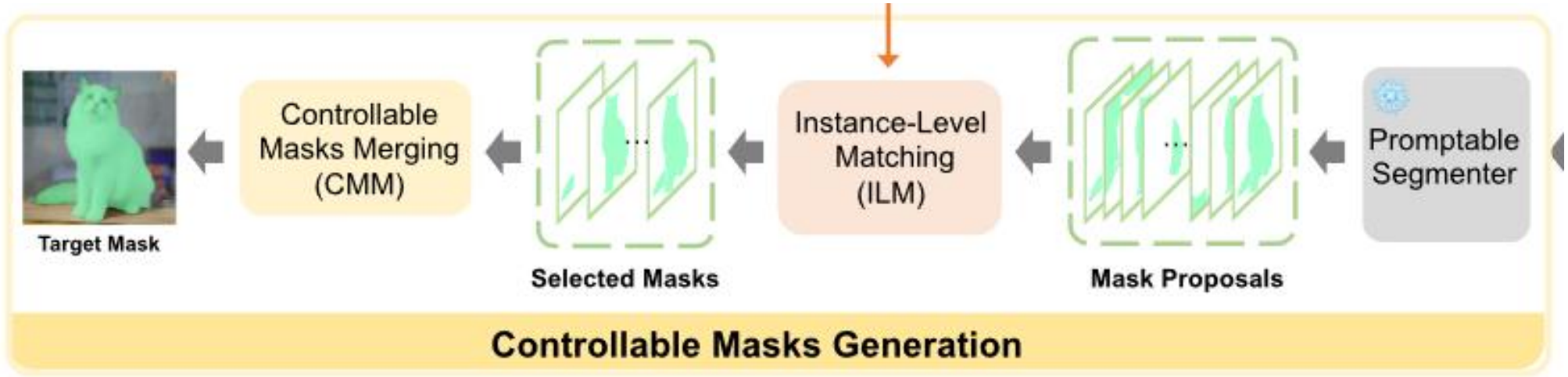
最后增加点的 Bounding Box 作为 Box Proposal

Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

研究现状-概念理解-自然语言-Matcher

Controllable Masks Generation (CMG)

- 选择高质量的Mask，然后合并选择的Masks以获得最终的目标Mask。
- mask 与 proposal 匹配：将匹配问题看做最优运输问题，衡量 mask 的相关性
- mask 质量选择：根据 mask 内关键点设置 coverage 与 purity 参数，衡量 mask 的质量



Yang Liu, et al. "Matcher: Segment anything with one shot using all-purpose feature matching.”, arXiv:2305.13310

03

总结与展望

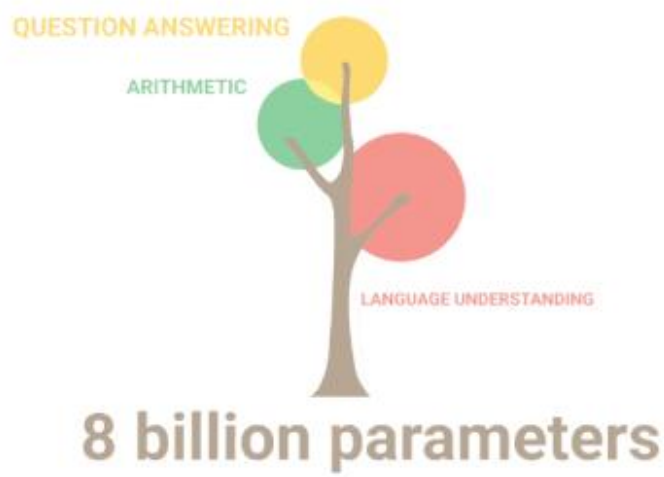
总结与展望-1

探索新的廉价可扩增数据，形成新的开放视觉能力

- **数据集规模越来越大：**从2015年的Flickr30k到2021年的CLIP数据集，其数据量规模从30k扩展到了400M级别，增长了1万多倍。近期发布的LAION-5B数据集甚至有5B的图像-文本对数据。
- **表达能力越来越强：**通过使用大规模数据进行训练，模型能够学习到更多语义知识，使其在处理新的、开放性的任务时具有更强的鲁棒性和适应性。
- 为了构建更好的开放视觉感知模型，需要探索类似于图像-文本对的**廉价可扩增数据模式**。

总结与展望-2

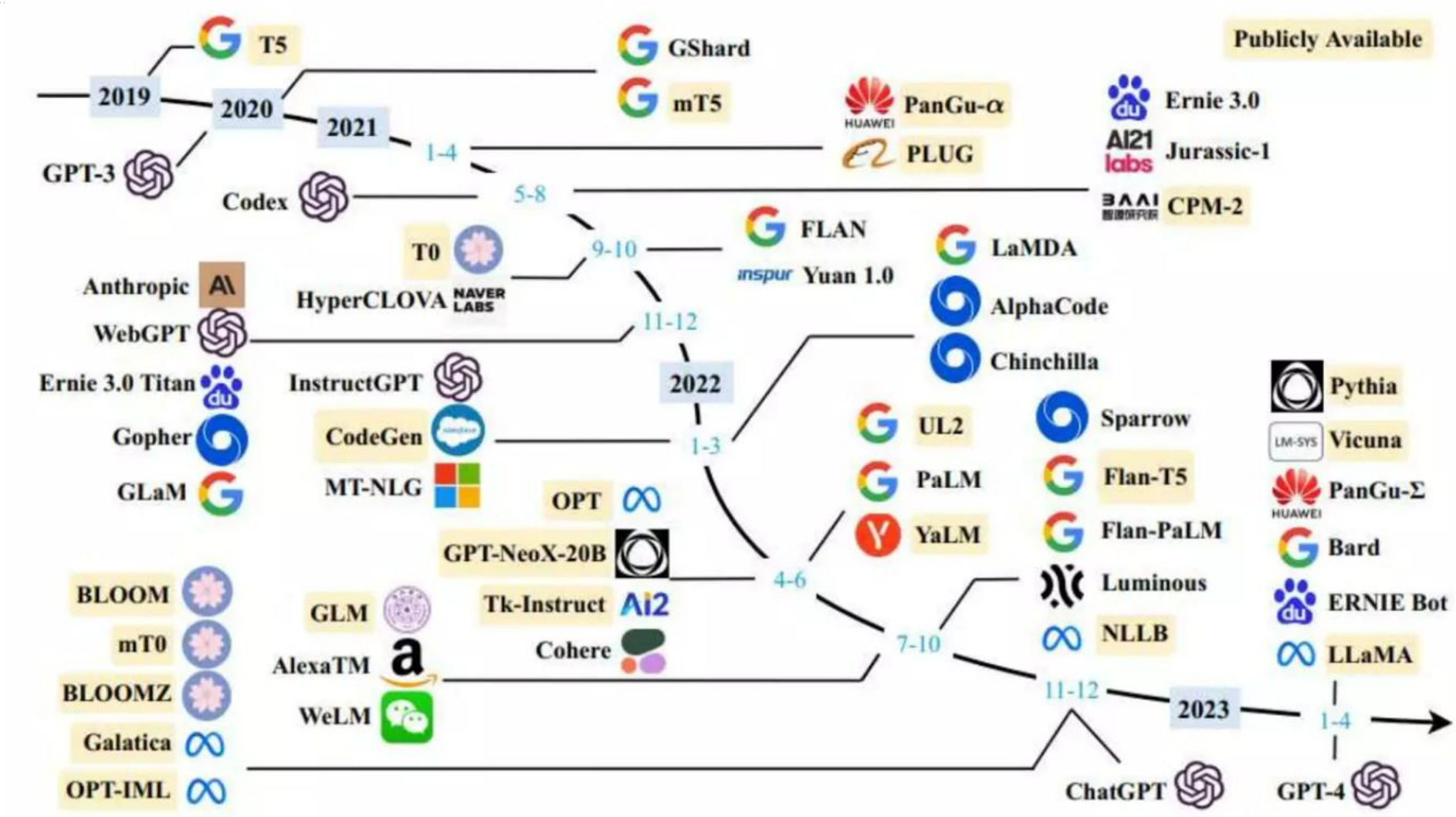
大语言模型涌现出的新能力能更好地理解新概念，赋能开放感知任务



大语言模型（LLMs）能力涌现现象：推理能力，in context learning

Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance; Sharan .et al.

LLM发展进程



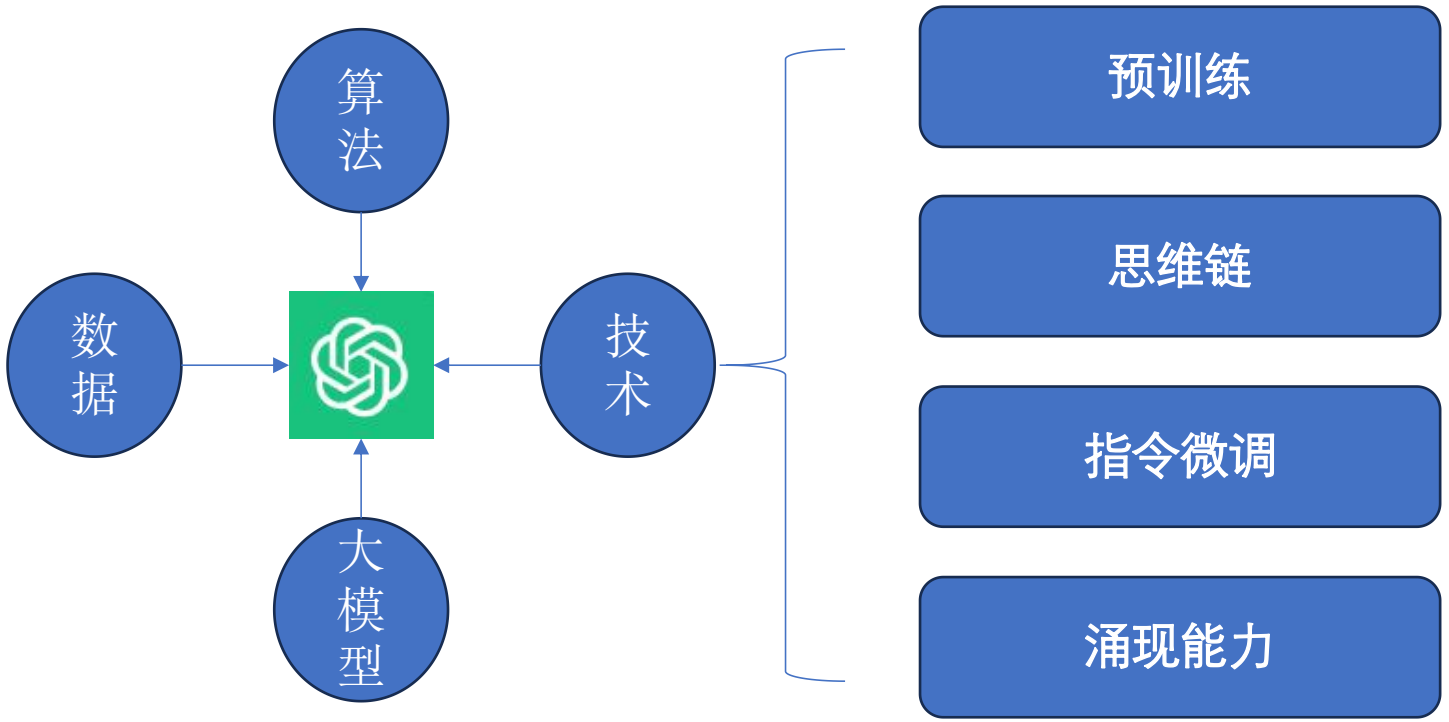
从2019年开始，大语言模型的发展迎来了一个新时代，其中大规模预训练模型成为主导，呈现了一片百花齐放的场面，其中主要以闭源的**ChatGPT系列**和开源的**LLaMA家族**为代表性工作

ChatGPT概述

ChatGPT是OpenAI开发的大型语言模型，它使用大规模自监督学习技术在大量文本数据上训练，通过理解和生成人类语言来回答问题的一种生成式人工智能技术。

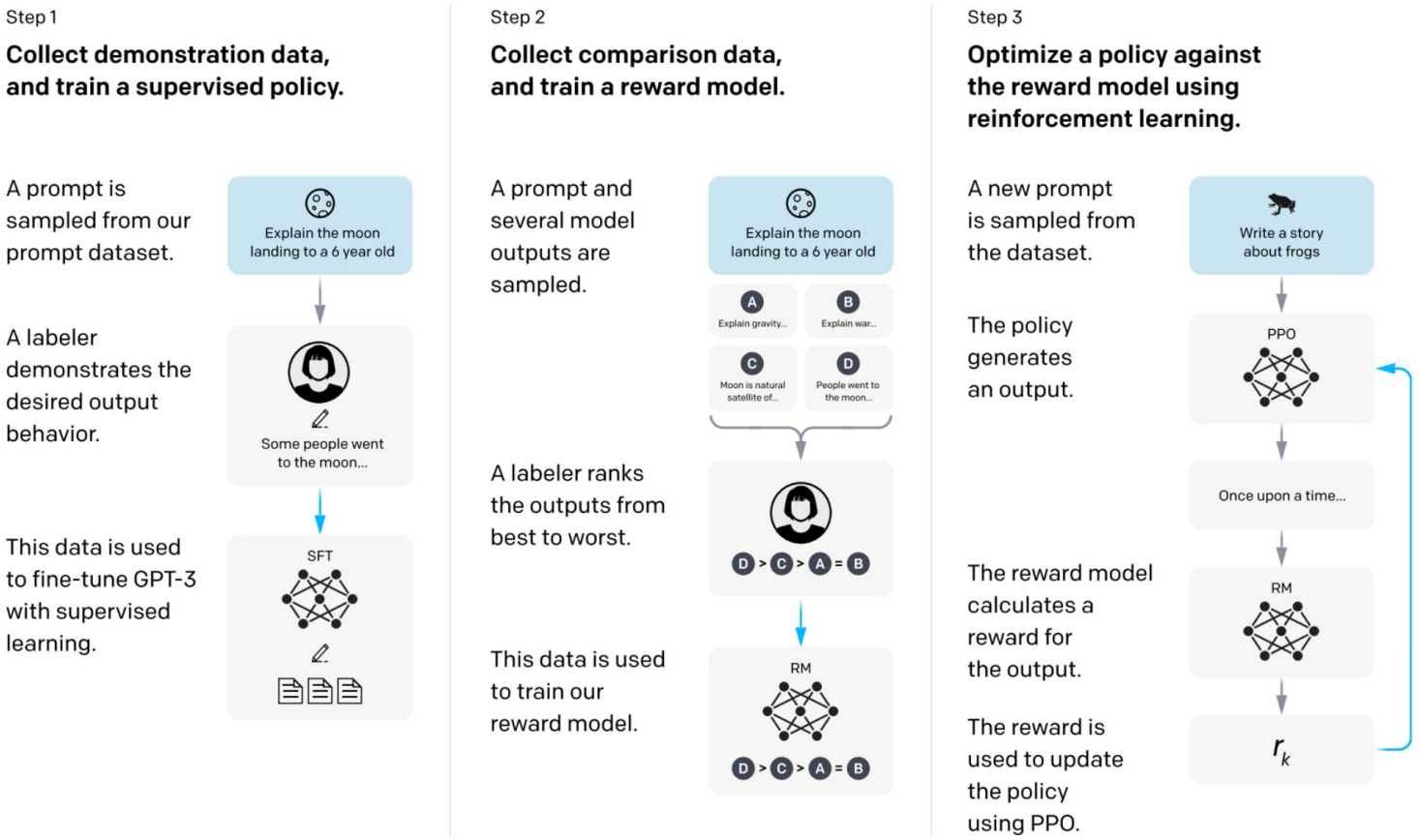
ChatGPT的核心技术包括：

- **大规模预训练**：保证模型以无监督模式充分学习大量数据，是各种能力的前提
- **思维链**：有助于模型理解人类思考的方式，以及在特定任务上的认知过程
- **指令微调**：对通用的预训练模型针对特定任务或指令进行优化和调整，以提高任务性能
- **涌现能力**：在模型训练的过程中，模型表现出的新的未曾预测或设计到的能力或特性。



ChatGPT训练过程

1. **无监督预训练：**直接使用大规模的文本数据自回归训练。
2. **监督微调：**引入多种具体的任务数据集训练模型预测答案。
3. **指令微调：**
    1. 从测试用户提交的问答中随机抽取数据，让专业的标注人员给出高质量的答案，并使用这些数据优化模型。
    2. 使用前面的模型生成N个不同的回答，让专业的标注人员对回答的质量进行排序，并使用这些数据训练一个奖励模型。
    3. 利用前面训练好的奖励模型，无需人工标注，通过强化学习的方式自动更新模型参数。



LLaMA训练数据

LLaMA训练数据包括网站、书籍、在线论文仓库、维基百科等。LLaMA 预训练数据大约包含 1.4T tokens，对于绝大部分的训练数据，在训练期间模型只见到过1次，Wikipedia 和 Books 这两个数据集见过2次。

右表所示是 LLaMA 预训练数据的含量和分布，其中包含了 CommonCrawl 和 Books 等不同域的数据。

对于不同的数据，作者还进行了不同的预处理。例如，使用去重和语言识别步骤，基于行长度或字母数字字符比例的启发式方法过滤低质量文件，并使用正则表达式删除了诸如头文件之类的样板文件等等

数据集	样本比例	Epochs	磁盘大小
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

LLaMA模型架构

LLaMA 的网络基于 Transformer 架构。作者利用了随后提出的各种改进，这些改进在不同模型（如 PaLM）中得到了应用。以下是与原始架构的主要区别，以及从哪里得到了这种变化的灵感。

Pre-normalization [受 GPT3 的启发]

为了提高训练稳定性，LLaMA 对每个 Transformer 子层的输入进行归一化，而不是对输出进行归一化。LLaMA 使用了 RMSNorm 归一化函数。

Rotary Embeddings [受 GPTNeo 的启发]：

LLaMa 没有使用之前的绝对位置编码，而是使用了旋转位置编码（RoPE），可以提升模型的外推性。

SwiGLU 激活函数 [受 PaLM 的启发]：

LLaMA 使用 SwiGLU 激活函数替换 ReLU 以提高性能。
SwiGLU 是2019年提出的新的激活函数，它结合了 SWISH 和 GLU 两种者的特点。SwiGLU 主要是为了提升Transformer 中的 FFN(feed-forward network) 层的实现。

LLaMA实验结果

在之前的研究的基础上，作者考虑了 zero-shot 和 few-shot，并在总共20个基准上报告了结果：作者将 LLaMa 与其他基础模型进行了比较，包括 GPT-3、Gopher、Chinchilla 和 PaLM，以及开源的 OPT 模型、GPT-J 和 GPT-Neo。

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

常识推理上的实验结果

		MATH	+maj1@k	GSM8k	+maj1@k
PaLM	8B	1.5	-	4.1	-
	62B	4.4	-	33.0	-
	540B	8.8	-	56.5	-
Minerva	8B	14.1	25.4	16.2	28.4
	62B	27.6	43.4	52.4	68.5
	540B	**33.6**	**50.3**	**68.5**	**78.5**
LLaMA	7B	2.9	6.9	11.0	18.1
	13B	3.9	8.8	17.8	29.3
	33B	7.1	15.2	35.6	53.1
	65B	10.6	20.5	50.9	69.7

在数学应用题 GSM8k 上的实验结果

总结与展望-2

构建视觉-语言统一表示，实现开放视觉感知

搜寻有价值的物体

特征统一表示

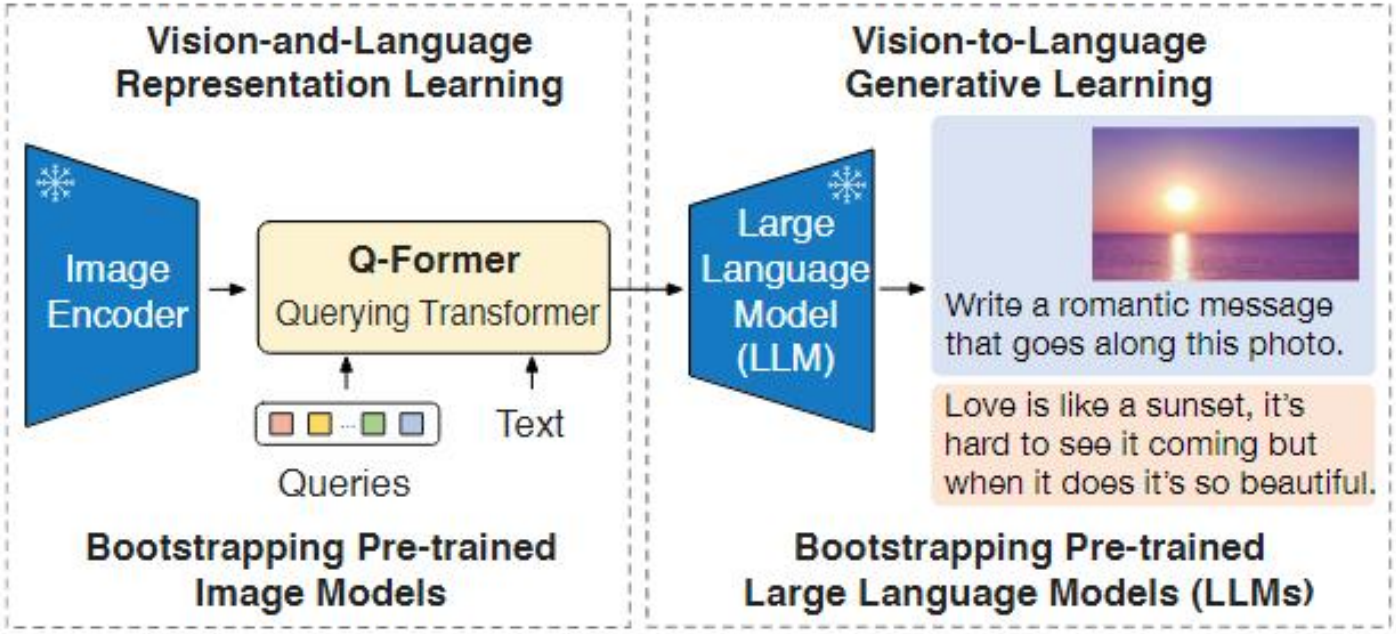
大语言模型

定位结果

映射视觉空间到语言空间，建立特征统一表示

通过大语言模型的推理能力，理解语言引申含义

BLIP-2

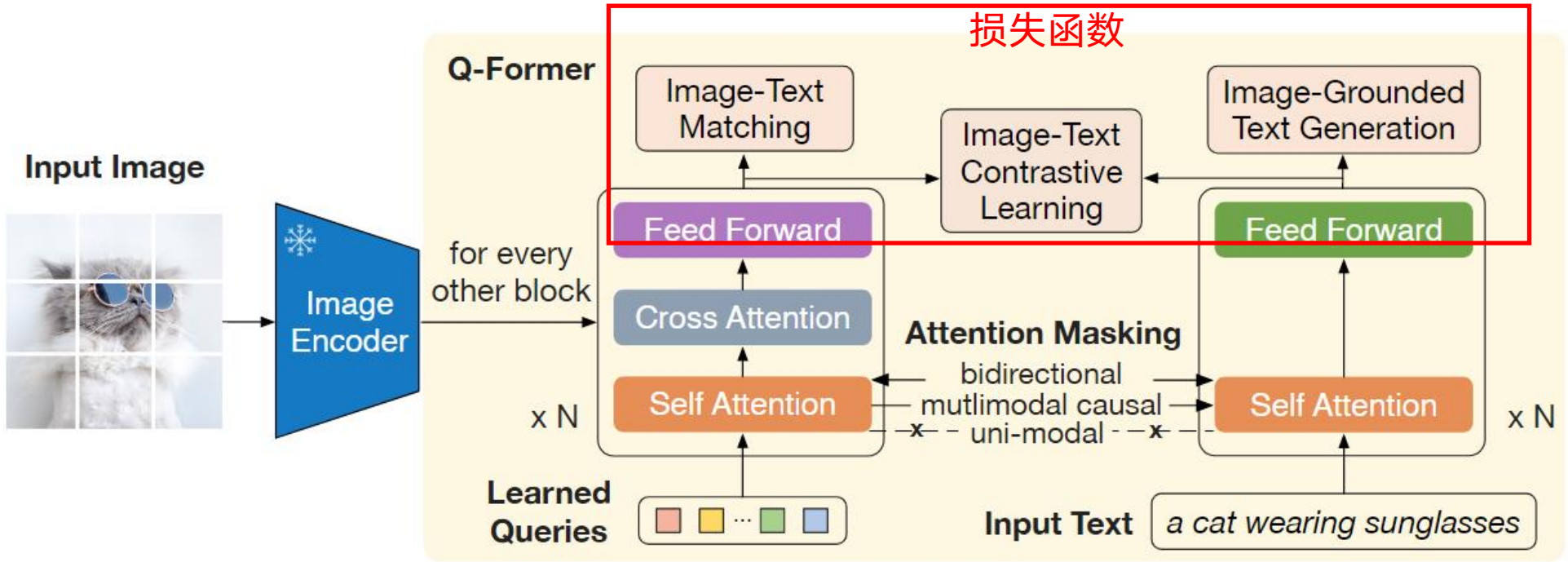


动机：视觉-语言预训练模型规模越来越大，由于使用了大模型、大数据集，而且采取端到端的训练，大多数最先进的视觉-语言模型在预训练过程中会产生很高的计算代价和经济成本

为了解决这个问题，BLIP-2 提出了一个轻量级的 Querying Transformer，通过一组**可学习的轻量 Query 特征**从冻结的视觉编码器中提取视觉特征，并充当视觉编码器和文本编码器之间的桥梁

Junnan Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." ICML 2023.

BLIP-2



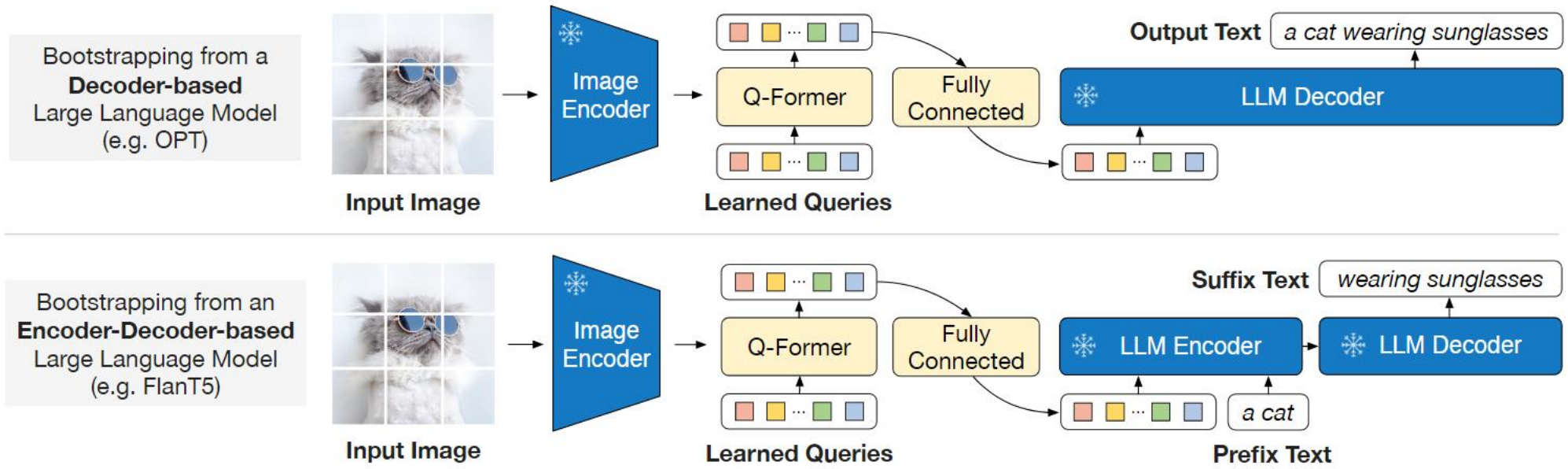
Q-Former的训练过程主要分为两阶段：在第一阶段

- **学习与文本最相关的视觉表征：**通过将 Q-Former 连接到冻结参数的图像编码器，基于Queries特征采样图像特征并使用图像-文本对进行预训练，从而使得 Queries 可以**学习到如何更好地结合文本提取图片信息**

- 通过上述训练过程可以有效地将最相关的视觉信息提供给 LLM，同时删除不相关的视觉信息

Junnan Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." ICML 2023.

BLIP-2



Q-Former的训练过程主要分为两阶段：在第二阶段

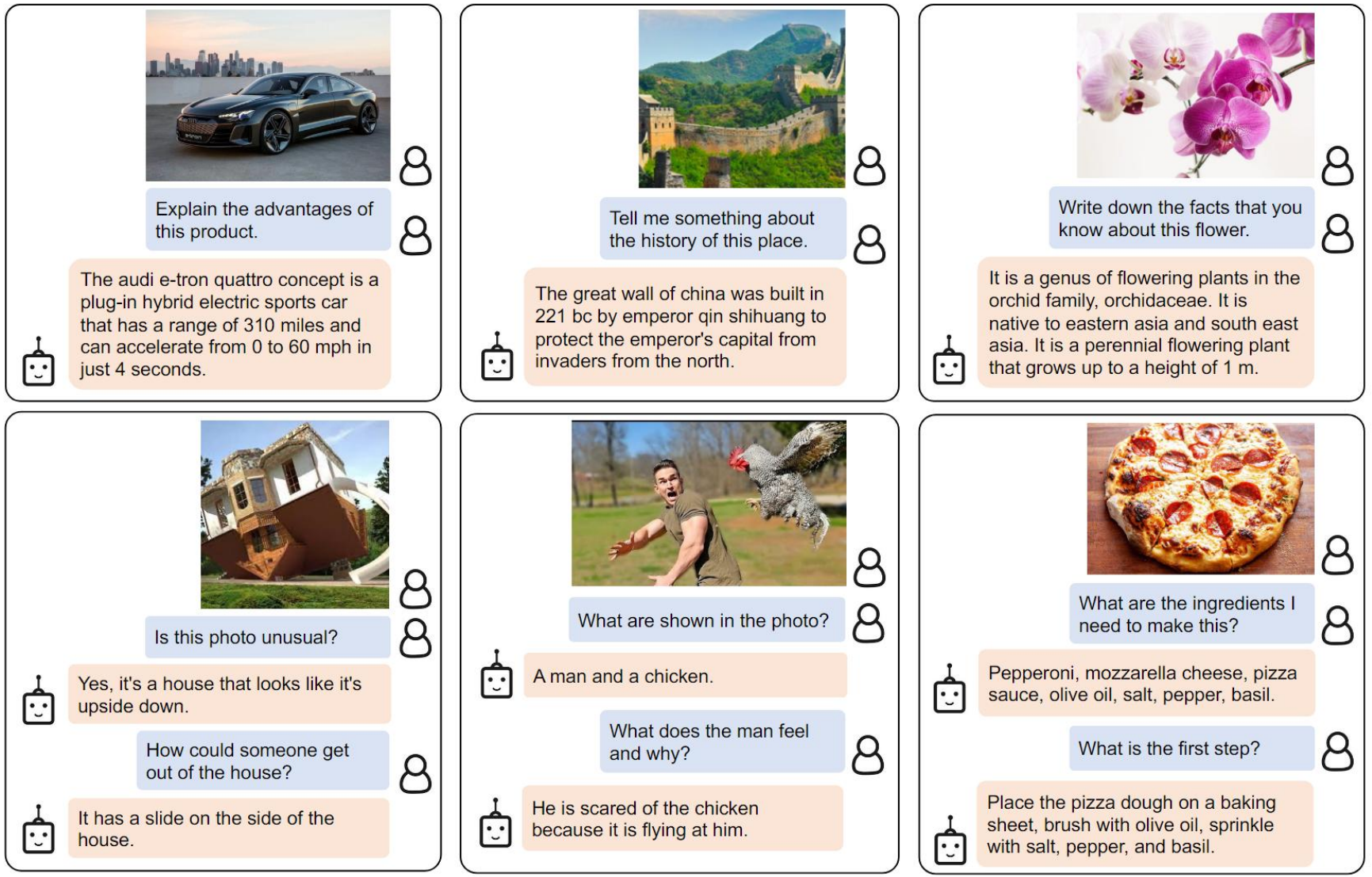
- **学习通过视觉特征重建文本特征：**
  - 对基于纯 Decoder 架构的模型，使用语言建模目标函数进行训练，冻结参数的 LLM 的任务是根据 Q-Former 提供的视觉表征来生成文本
  - 对基于 Encoder-Decoder 架构的模型，把文本分成两段，前缀随着 Queries 的输出送入 LLM 的 Encoder，希望 Decoder 输出后缀文本信息

- 通过上述训练过程可以实现视觉特征向语言大模型特征表示的适配

Junnan Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." ICML 2023.

BLIP-2

通过上述方式可以实现**让LLM理解视觉输入**



Junnan Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." ICML 2023.

KOSMOS-2

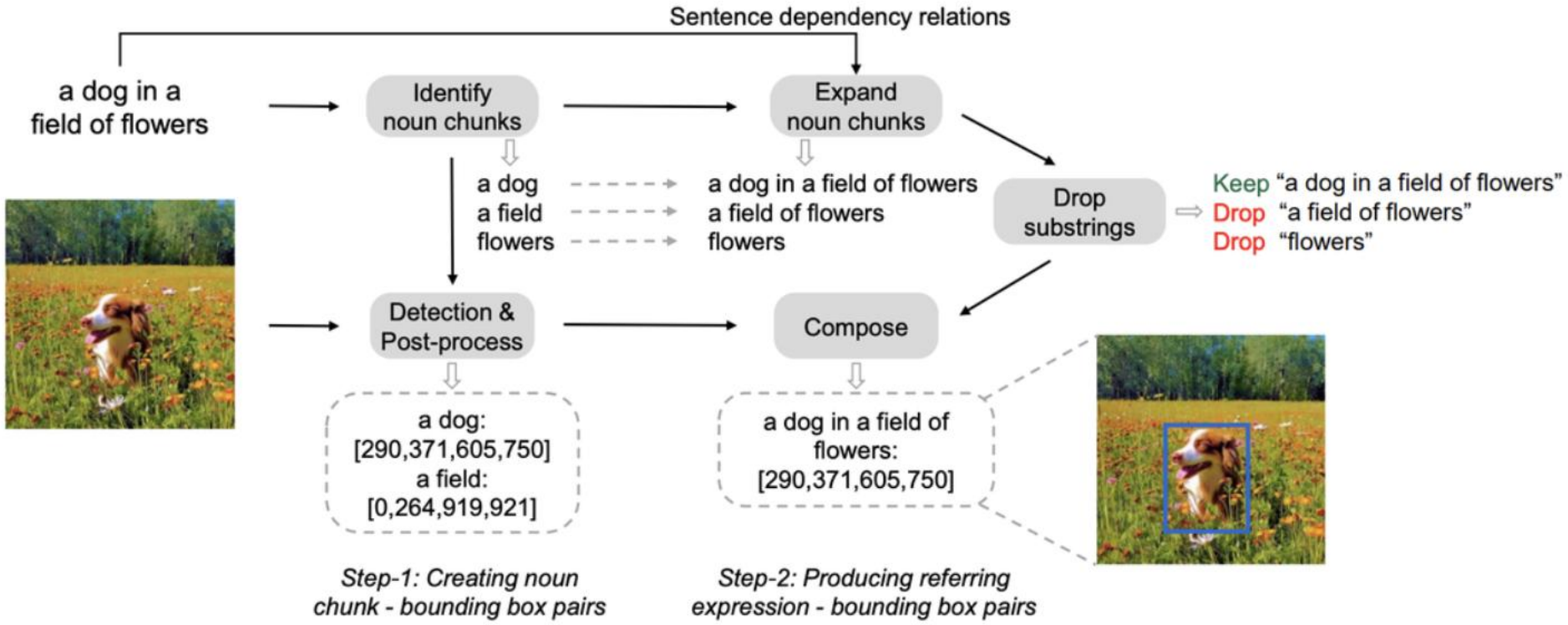
KOSMOS-2解锁了多模态大语言模型的**定位能力**。定位能力可以为视觉-语言任务提供更方便，更有效的人机交互。它使用户能够直接指出图像中的目标或区域，而不是输入详细的文本说明来引用它，模型可以理解具有空间位置的图像区域

- 以**超链接**的方式引入物体的位置信息
- 以**离散化**的position token编码连续位置信息

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

KOSMOS-2

包含位置信息的数据构造过程



- 利用SpaCy识别句子中的名词，与图片一同输入到GLIP中，得到匹配结果
- 为了保证名词的复杂性，利用SpaCy生成的语法树对名词进行扩展，删除被其他句子包含的名字，得到最终的text-box pair
- 上亿的数据量

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

KOSMOS-2

位置信息引入

<s> <image> Image Embedding </image> <grounding> It <box><loc$_{44}$><loc$_{863}$></box>
seats next to a campfire <box><loc$_{4}$><loc$_{1007}$></box> </s>

- 使用<box></box>这种的特殊token识别bounding box
- 将一张图片划分为pxp的格子，每个格子对应着一个special token，所有落在格子里的像素都被指定为这个格子，例如<loc44>，那么一个box的左上角坐标和右下角坐标被转换为两个token，与传统的方法相比token数量少了几乎10倍
- 对于一个名词指代多个box的情况，使用</delim>作为分隔符

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

KOSMOS-2

训练过程

阶段一：类似kosmos-1中的方法，使用多模态语料库进行模型预训练

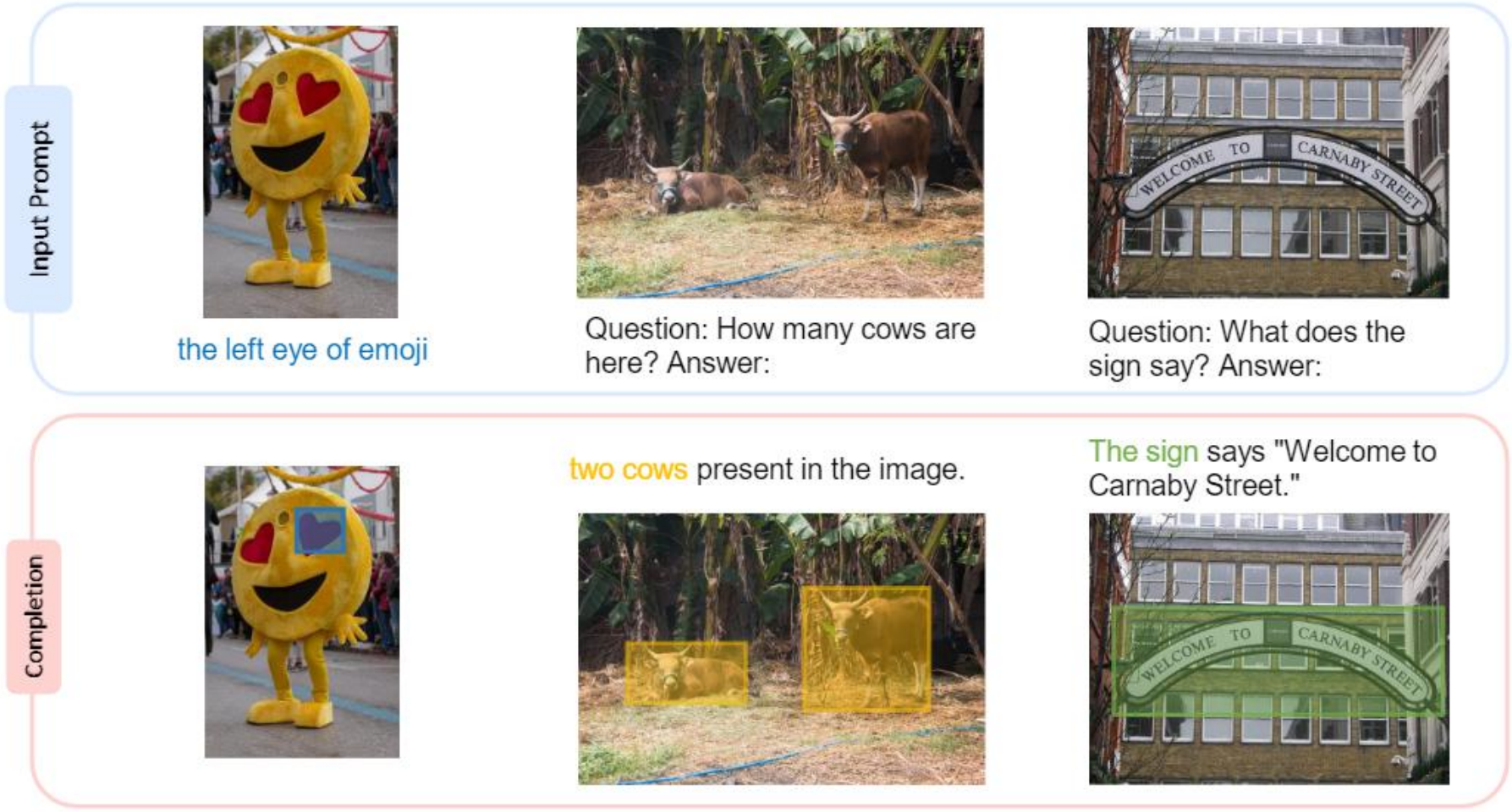
阶段二：指令微调，数据使用了

Llava中的vision-language instruction dataset
Language-only instruction datasets
GRIT数据，转换为Instruct tuning的结构
GRIT数据，利用下面的模版转换为captioning任务

- "What is it <box><loc$_1$><loc$_2$></box>? It is {*expression*}."
- "What is this <box><loc$_1$><loc$_2$></box>? This is {*expression*}."
- "Describe this object <box><loc$_1$><loc$_2$></box>. This object is {*expression*}."
- " It <box><loc$_1$><loc$_2$></box> is {*expression*}."
- " This <box><loc$_1$><loc$_2$></box> is {*expression*}."
- " The object <box><loc$_1$><loc$_2$></box> is {*expression*}."

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

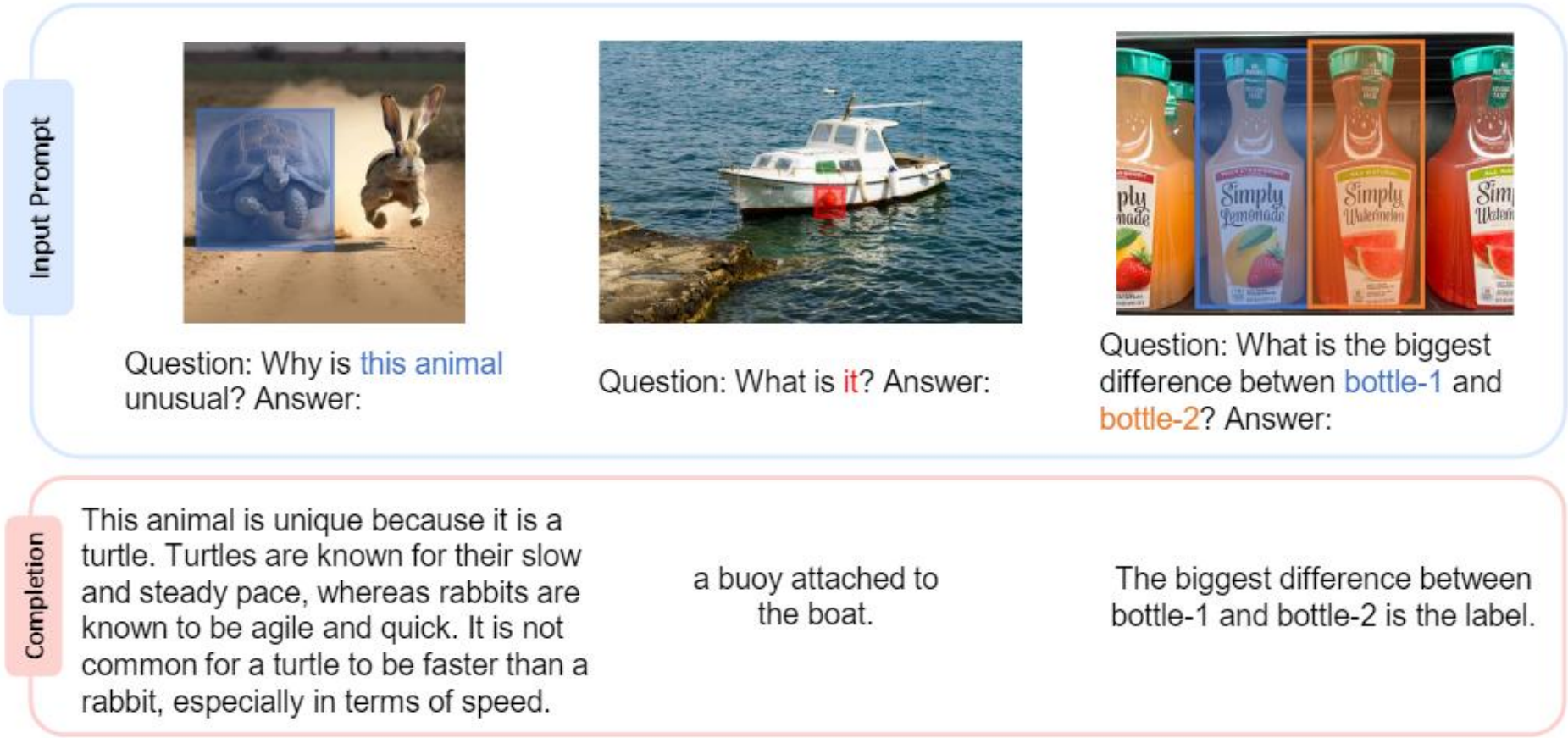
KOSMOS-2



视觉定位与定位问答

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

KOSMOS-2



多模态边界框指代分析

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

KOSMOS-2

Input Prompt

Describe this image in detail:

Completion

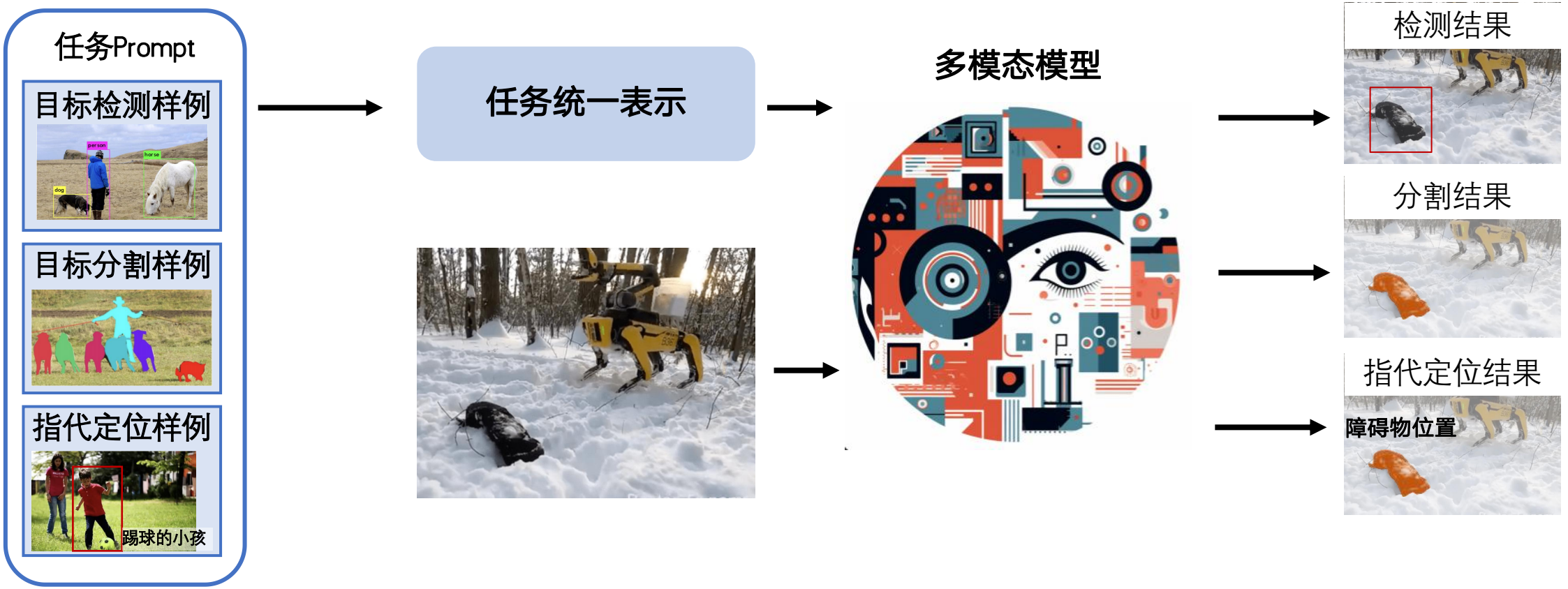
A snowman sits next to a campfire in the snow. He is wearing a hat, scarf, and mittens. There are several pots nearby, likely containing a hot meal for the snowman. The scene is serene and peaceful, with the snowman looking at the fire with a warm smile.

定位图像字幕

Zhiliang Peng et al. "Kosmos-2: Grounding Multimodal Large Language Models to the World.”, arXiv:2306.14824

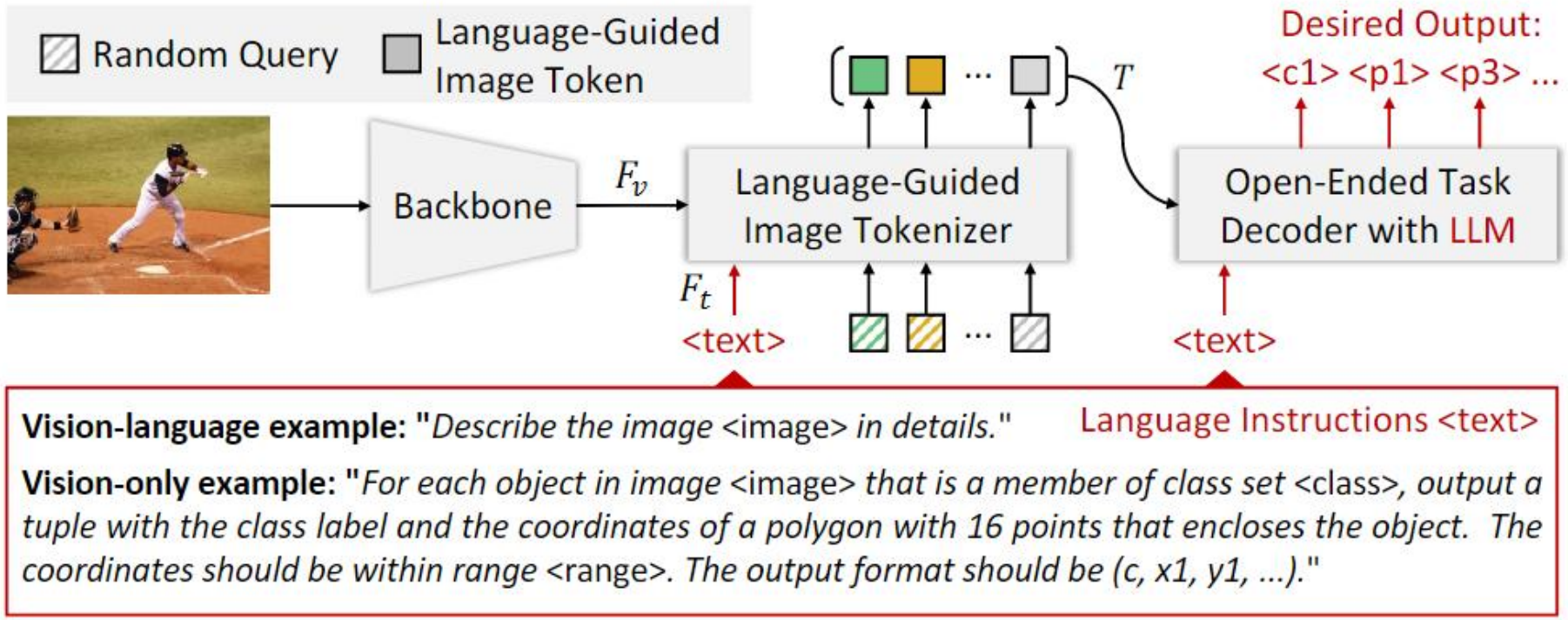
总结与展望-2

构建多视觉任务的统一表示，实现通用视觉感知模型



通过多模态大模型 in context learning 能力，解决自定义的视觉感知任务

VisionLLM

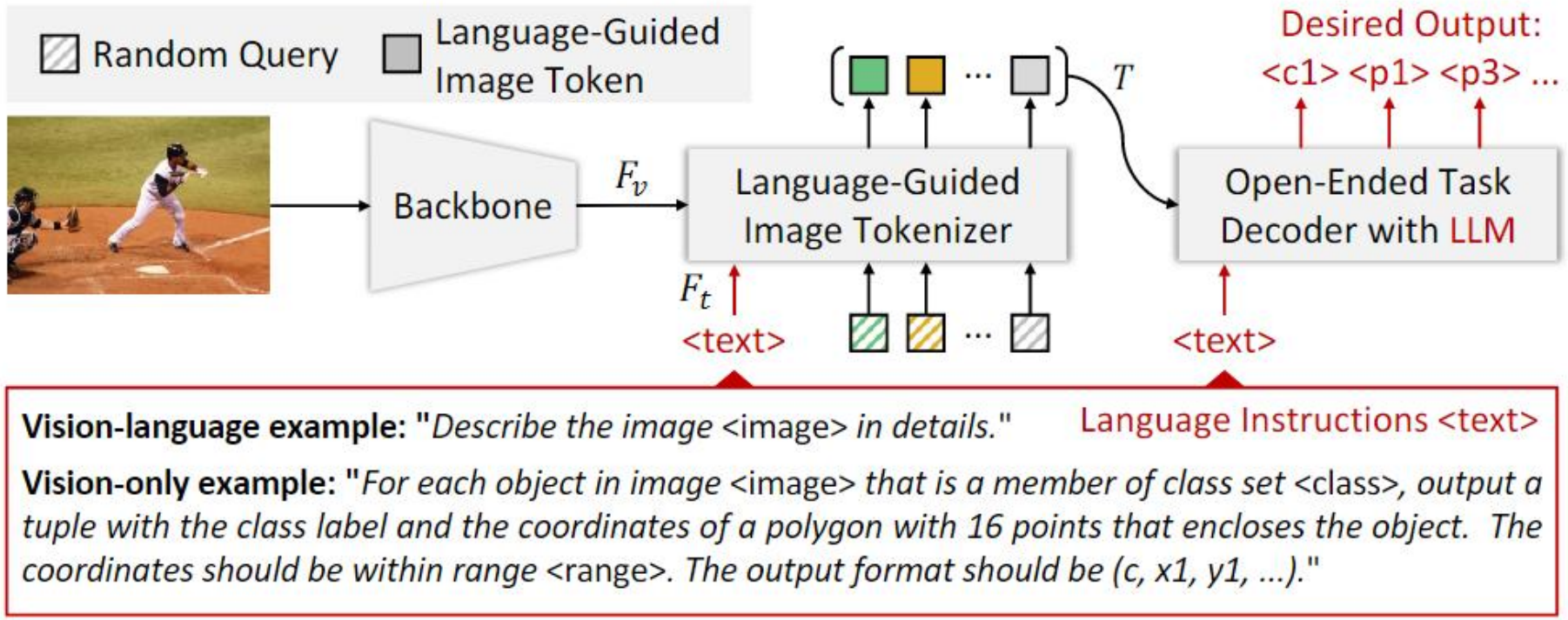


动机：尽管有众多强大的视觉基础模型（VFMs）可用，但它们仍然限制于预定义形式的任务，难以匹配LLMs的开放式任务能力

提出了一种基于LLM的**视觉中心任务框架**，该框架通过将图像视为外语，并将视觉中心任务与可以使用语言指令灵活定义和管理的语言任务进行对齐，为视觉和语言任务提供了**统一**的视角

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.”, arXiv:2305.11175

VisionLLM



整体框架组成：

1. 统一的语言指令模块，给视觉任务提供一致的接口
2. 语言引导的图像标记器，根据给定的语言提示对图像进行编码，使模型有效理解视觉内容
3. LLM引导的开放任务解码器，根据解码的视觉信息和语言指令得到预测输出

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.”, arXiv:2305.11175

VisionLLM

语言指令模块

视觉-语言任务

"The image is <image>. Please generate a caption for the image: "

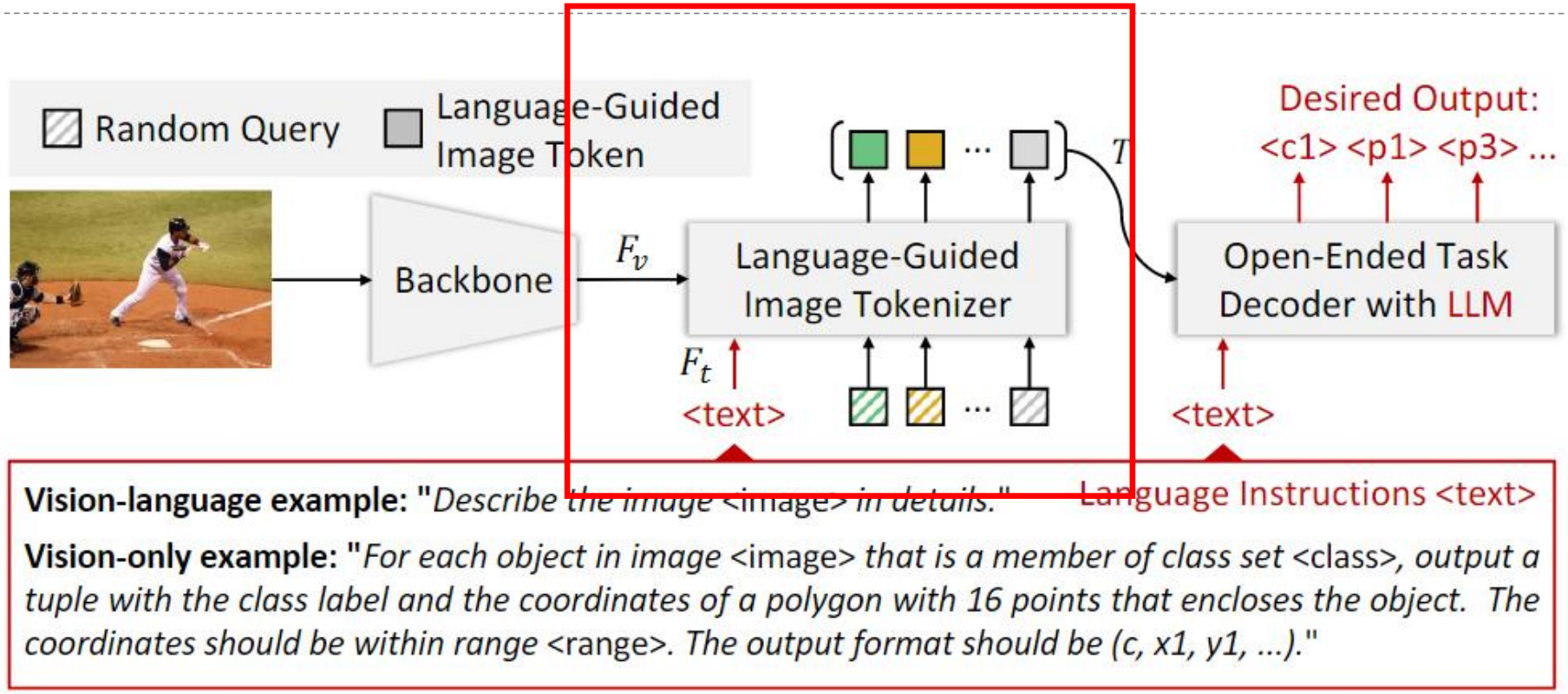
"The image is <image>. Please generate an answer for the image according to the question: <question>"

纯视觉任务

An example of language instruction for the instance segmentation task is as follows: "*Segment all the objects of category set* `<class>` *within the* `<range>` *of the image and generate a list of the format (c, x1, y1, x2, y2, ..., x8, y8). Here, c represents the index of the class label starting from 0, and (x1, y1, x2, y2, ..., x8, y8) correspond to the offsets of boundary points of the object relative to the center point. The image is:* `<image>`".

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.", arXiv:2305.11175

VisionLLM

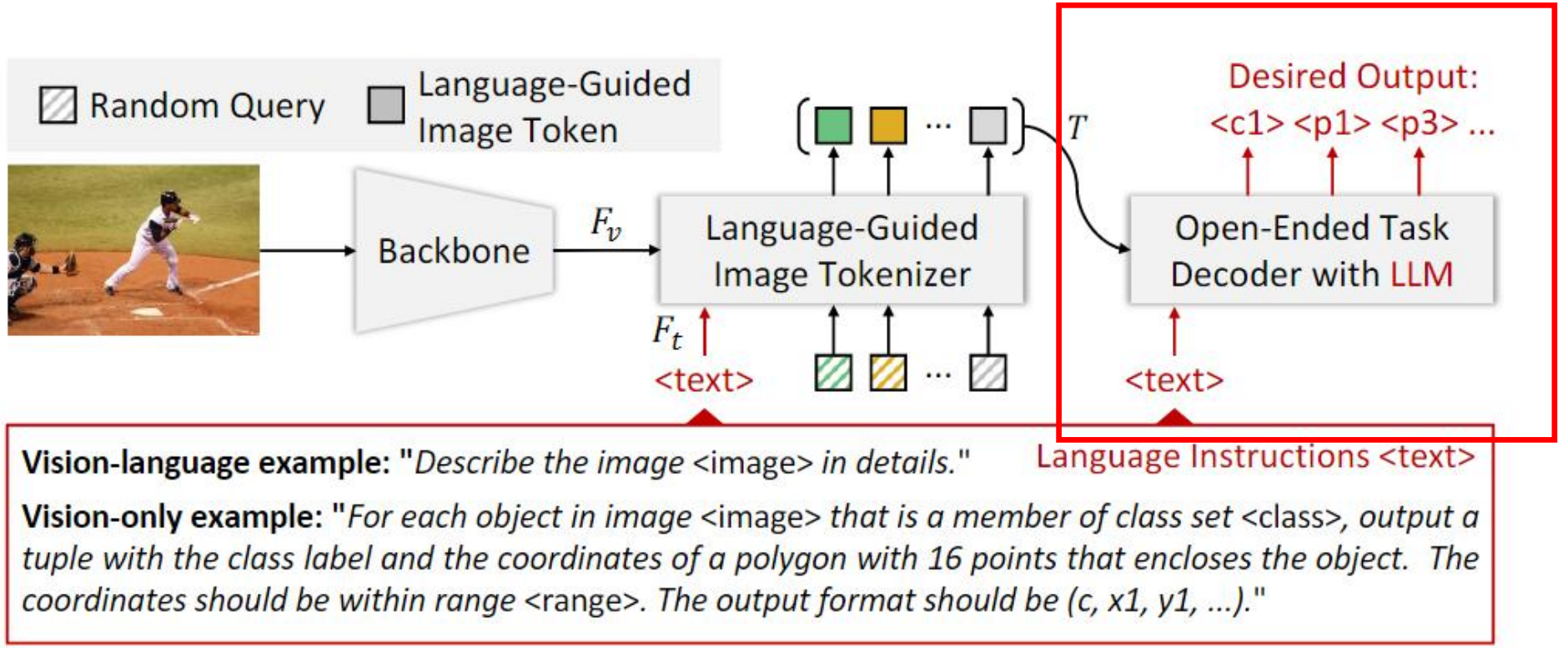


语言引导的图像标记器

多尺度图像特征与文本编码通过Cross-Attn融合，用一组可学习的Query与融合语义信息后的视觉特征交互，得到一系列输出为Image Token

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.”, arXiv:2305.11175

VisionLLM



LLM引导的开放任务解码器

基本框架：Alpaca

改进：构造了一组位置token：<p-512>等；构造了语义无关的类token<c1>等；

将问题中描述的格式作为Query（Output-Format-as-Query）

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.”, arXiv:2305.11175

VisionLLM

Human: "*Identify the objects in* <image> *that belong to {'What is the child eating?':* <c0>, *'red gamepad':* <c1>*} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where c is the class label and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box. The coordinates should be within the range 1024.*"

VisionLLM: "*The bounding boxes are [(*<c0>*, 226.4, 229.8, 363.1, 347.4), (*<c1>*, 441.1, 183.5, 538.6, 269.9)].*"

(a) Object Detection with *Customized Class Set (e.g., question, reasoning text, open-vocabulary description)*

Human: "*For each object in* <image> *that is a member of {'motorbike':* <c0>*}, output a tuple with the class label and the coordinates of a polygon with 16 points that encloses the object. The coordinates should be within the range 1024. The output format should be (c, x1, y1, x2, y2, ..., x16, y16).*"

VisionLLM: "*The detected polygons are [(*<c0>*, 135.3, 95.7, 123.4, 53.4, 84.9, 57.6, 66.8, 60.5, 60.1, 72.3, 34.2, 71.4, ..., 124.9, 119.3)].*"

(b) Object Location with *Editable Output Format*

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.", arXiv:2305.11175

VisionLLM

Human: "*Provide a long description of the* <image>."

VisionLLM: "*The image features a white dog laying down on a foldable outdoor cooling pad , which is placed on top of a wooden table. The dog has its head in the middle of the cooling pad, and a fan is positioned behind it, blowing air onto the pad. The table is outdoors, in a grassy area, and there are two chairs placed around it, one on each side. The scene appears to be set in a sunny weather, with a bright and sunny sky, which contributes to the overall outdoor atmosphere.*"

(c) Image Description with *Controllable Text Length*

Human: "*With the objects in the* <image>, *if you feel hot, what will you do?*"

VisionLLM: "*First, You can also try turning off the computer and other electrical appliances in the room to reduce the heat generated by their operation. As an alternative, you can use the desk fan provided in the image to create a cooling effect in the room. You can plug in the fan and adjust its settings to ensure it blows air in a direction that reaches the room's occupants effectively.*"

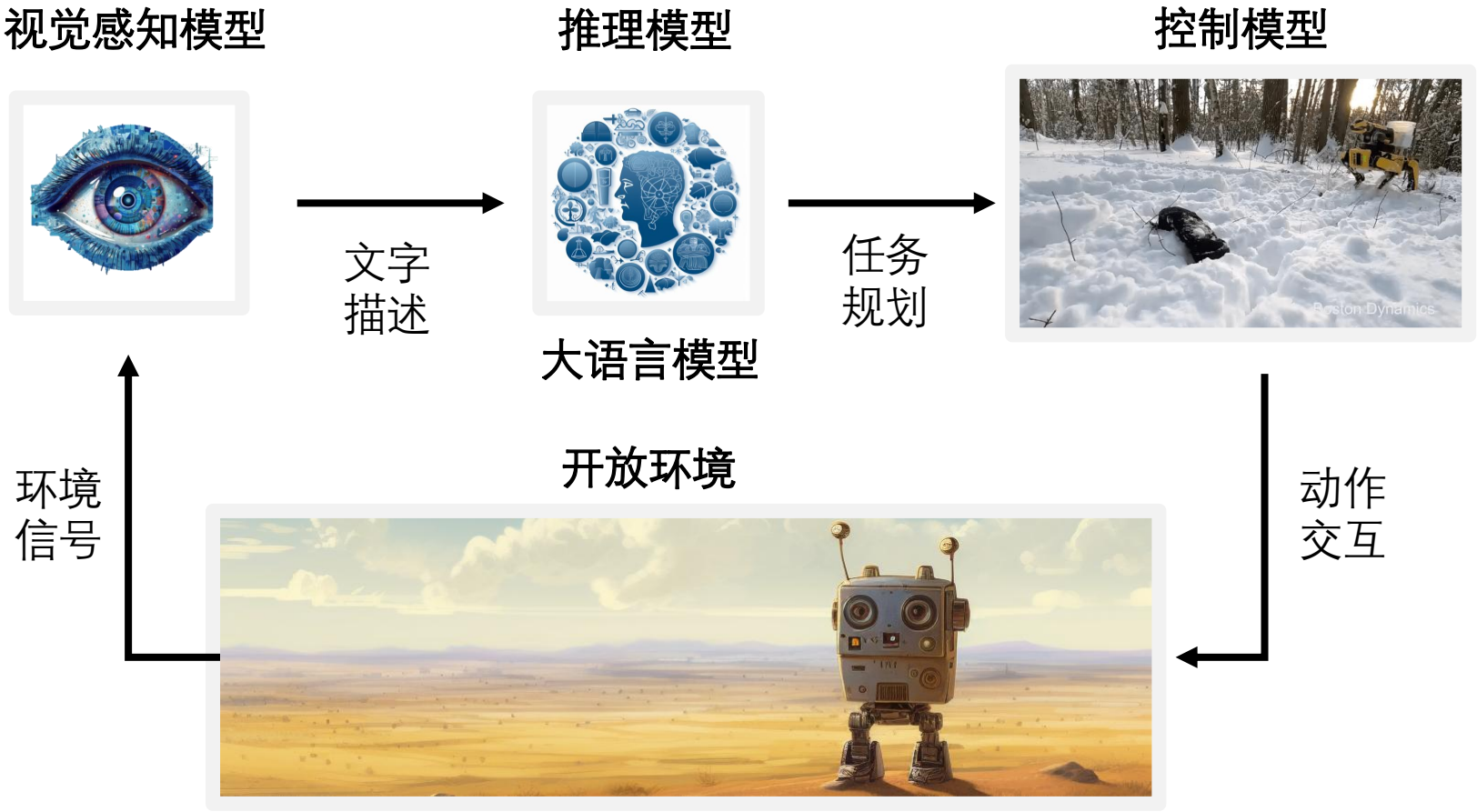
(d) Visual Question Answer with *Complex Reasoning*

Zhiliang Peng et al. "VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks.", arXiv:2305.11175

总结与展望-3

开放具身系统：开放感知系统的延伸　　被动感知->主动交互，感知智能->决策智能

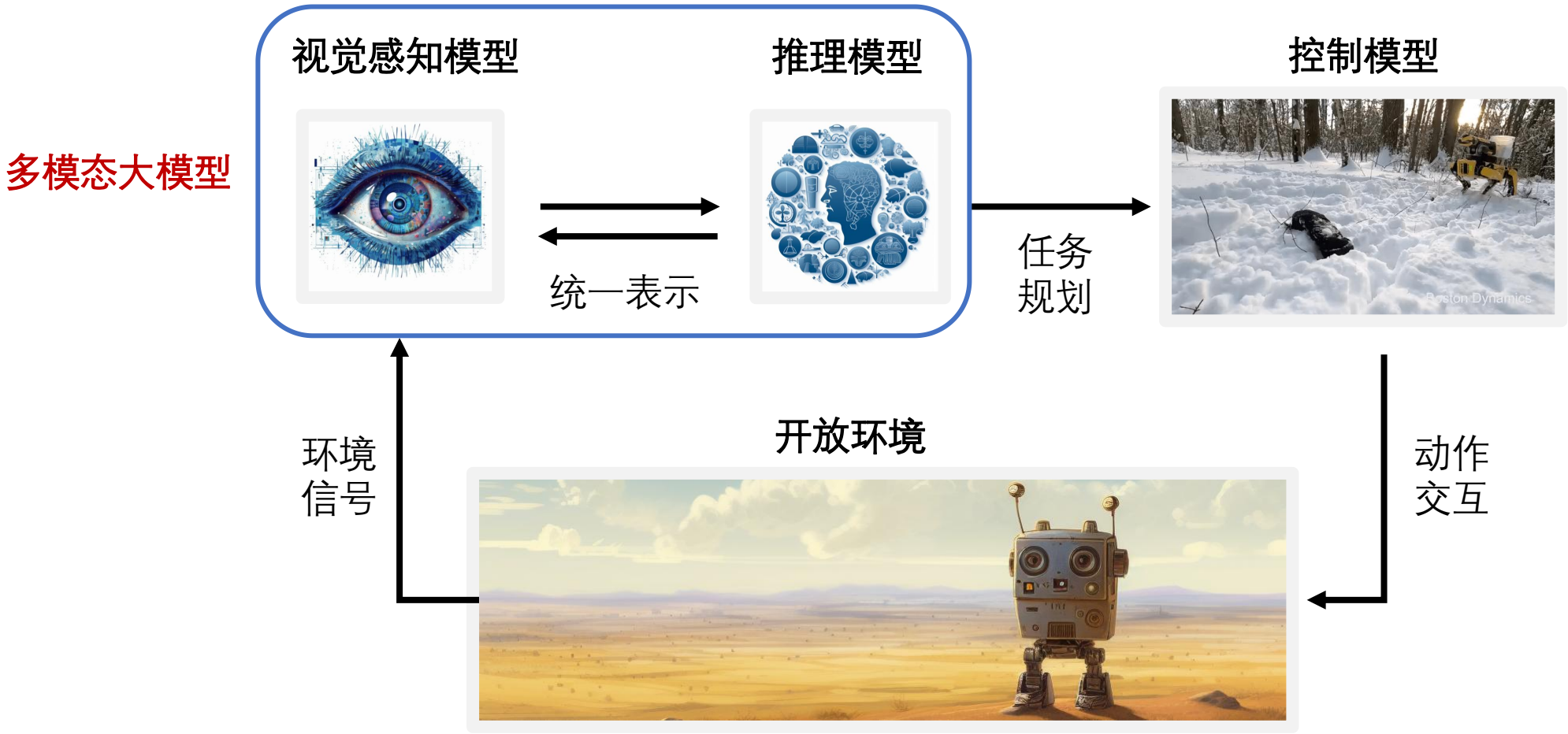
阶段一：以自然语言为媒介，实现对感知、推理、控制模型的协同调度



总结与展望-3

开放具身系统：开放感知系统的延伸　　被动感知->主动交互，感知智能->决策智能

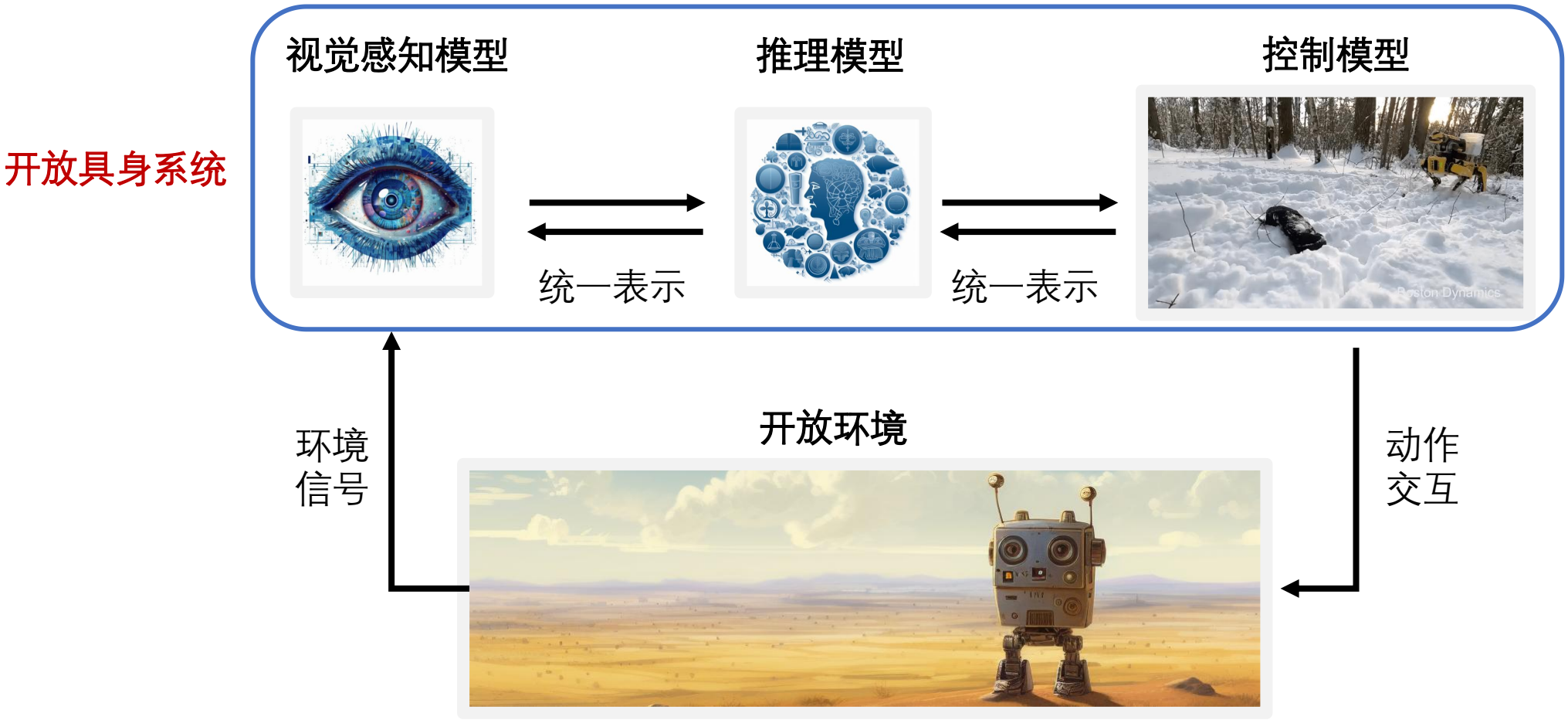
阶段二：通过构建“视觉-推理”统一表示，联合形成多模态大模型



总结与展望-3

开放具身系统：开放感知系统的延伸　　被动感知->主动交互，感知智能->决策智能

阶段三：统一“感知-推理-控制”特征空间

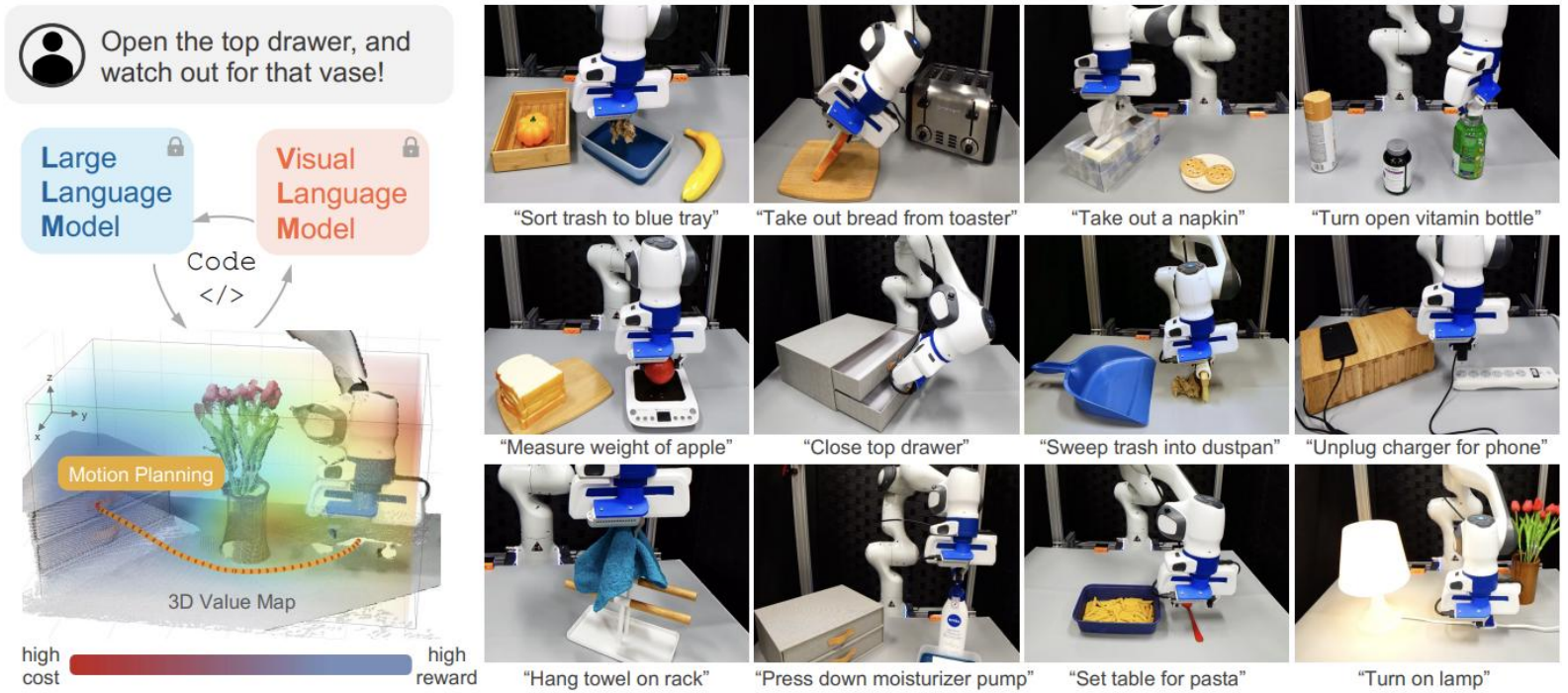


总结与展望-3

开放具身社区：构建一个虚拟环境，将众多的具身智能体放入其中，通过交互数据持续提升智能体能力

智能体自主与环境、其他智能体、以及人的数字孪生进行交互

VoxPoser

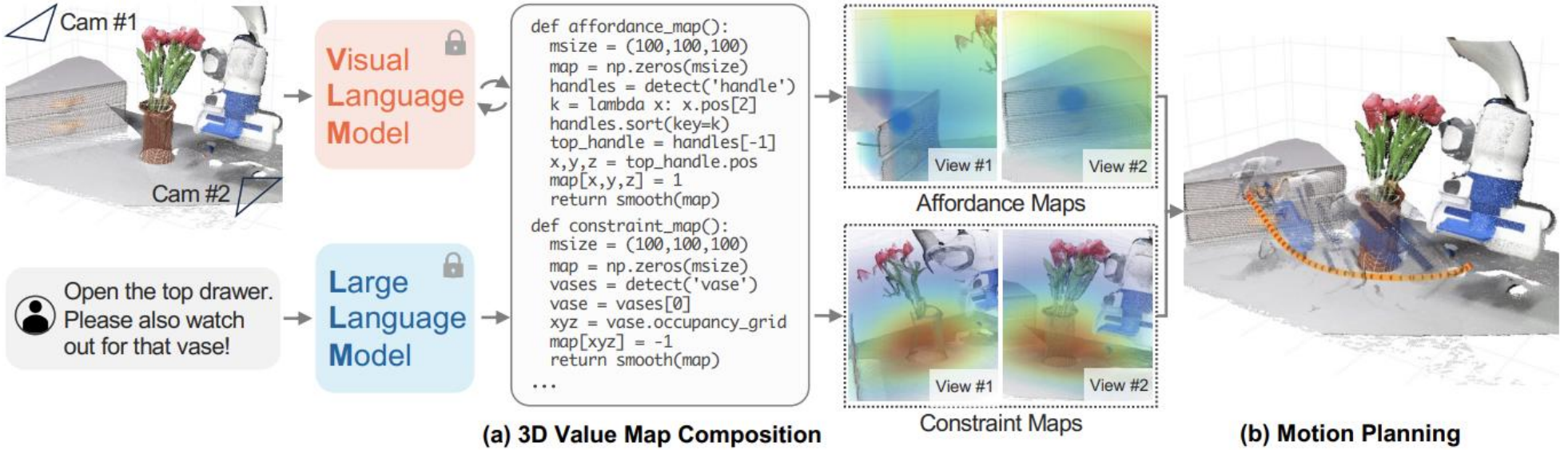


VoxPoser探索了大模型在机器人领域的应用：

- **大语言模型（LLMs）**在机器人操作中的应用，发现大语言模型可以根据任务输出相应的代码，生成完成所需任务的供应和约束。
- **视觉语言模型（VLM）**提供了代码与环境的交互的能力，通过与大预言模型的交互构建优化的3D价值地图。

Wenlong Huang et al. "VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models." CoRL 2023

VoxPoser



算法流程

1. **生成代码：**在给定环境的RGB-D观测和自然语言指令的情况下，大型语言模型（LLMs）生成代码。
2. **生成代价地图：**代码与视觉语言模型（VLMs）进行交互，以生成一系列3D可行性地图和约束地图
3. **轨迹规划：**这些组合的价值地图用作运动规划器的客观函数，用于合成机器人操作的轨迹

Wenlong Huang et al. "VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models." CoRL 2023

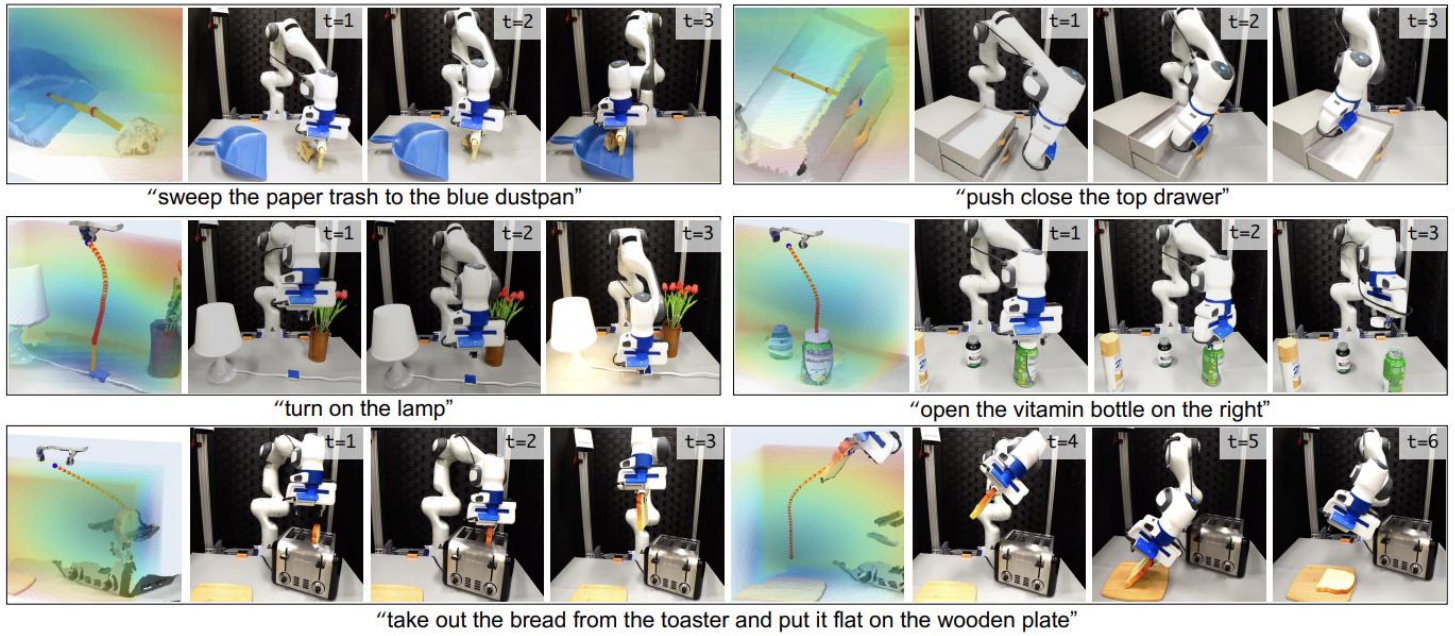
VoxPoser

实验结果

- VoxPoser的平均成功率较高。它能够成功执行各种日常操作任务，尤其在面对外部干扰时，表现出比基线模型更**灵活和鲁棒**的特点。
- VoxPoser在两个类别（共13项任务）中表现优于baseline，不仅在已见任务上表现出色，而且在未见任务上也具有相似的成功率，**在开放世界的环境下表现良好**
- 底部实验展现了语言模型和视觉模型的实际交互过程和效果

Task	LLM + Prim. [74]		VoxPoser	
	Static	Dist.	Static	Dist.
Move & Avoid	0/10	0/10	9/10	8/10
Set Up Table	7/10	0/10	9/10	7/10
Close Drawer	0/10	0/10	10/10	7/10
Open Bottle	5/10	0/10	7/10	5/10
Sweep Trash	0/10	0/10	9/10	8/10
Total	24.0%	14.0%	88.0%	70.0%

Train/Test	Category	U-Net	Language Models	
		MP [50]	Prim. [74]	MP (Ours)
SI SA	Object Int.	21.0%	41.0%	64.0%
SI SA	Composition	53.8%	43.8%	77.5%
SI UA	Object Int.	3.0%	46.0%	60.0%
SI UA	Composition	3.8%	25.0%	58.8%
UI UA	Object Int.	0.0%	17.5%	65.0%
UI UA	Composition	0.0%	25.0%	76.7%



Wenlong Huang et al. "VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models." CoRL 2023

北京航空航天大學
BEIHANG UNIVERSITY

Beihang University

谢谢大家，请批评指正

北京航空航天大学
刘偲
liusi@buaa.edu.cn