F-VLM: OPEN-VOCABULARY OBJECT DETECTION UPON FROZEN VISION AND LANGUAGE MODELS

ICLR 2023

报告人：徐静远

目录

目录

作者介绍

Weicheng Kuo*, Yin Cui†, Xiuye Gu†, AJ Piergiovanni*, Anelia Angelova*
*Google Research, Brain Team; †Google Research, Perception
{weicheng, yincui, xiuyegu, ajpiergi, anelia}@google.com

Weicheng Kuo

FOLLOW

Google Brain
Verified email at google.com - Homepage
Computer Vision

TITLE	CITED BY	YEAR
Deepbox: Learning objectness with convolutional networks W Kuo, B Hariharan, J Malik Proceedings of the IEEE international conference on computer vision, 2479-2487	204	2015
Open-vocabulary object detection via vision and language knowledge distillation X Gu, TY Lin, W Kuo, Y Cui arXiv preprint arXiv:2104.13921	166	2021
Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning W Kuo, C Häne, P Mukherjee, J Malik, EL Yuh Proceedings of the National Academy of Sciences 116 (45), 22737-22745	163	2019

目录

研究背景一：视觉语言模型

- 可用于开放类的视觉语言模型
  - 以CLIP 为例[1]

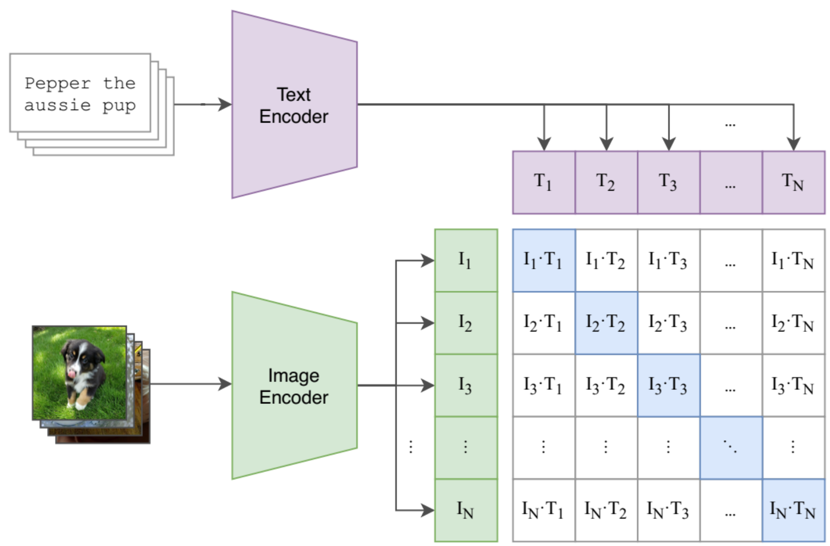


图1：CLIP模型的训练方式，4亿训练对

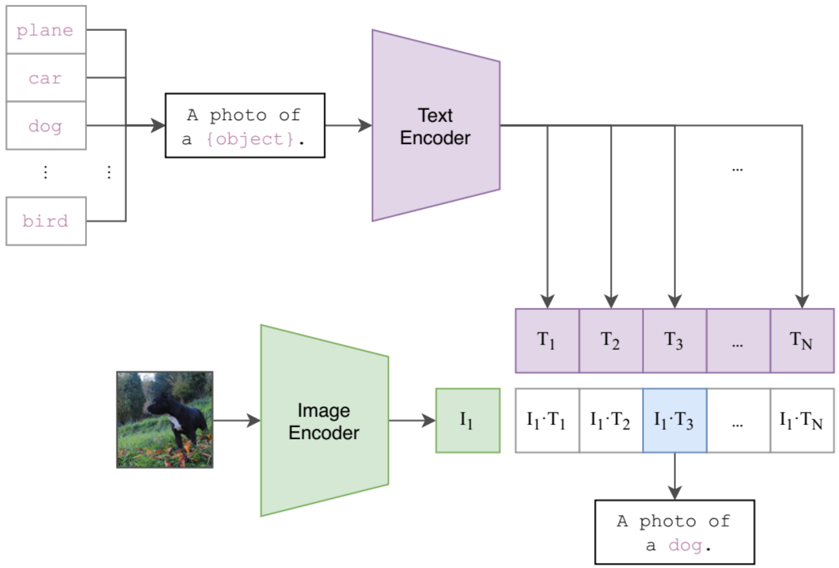
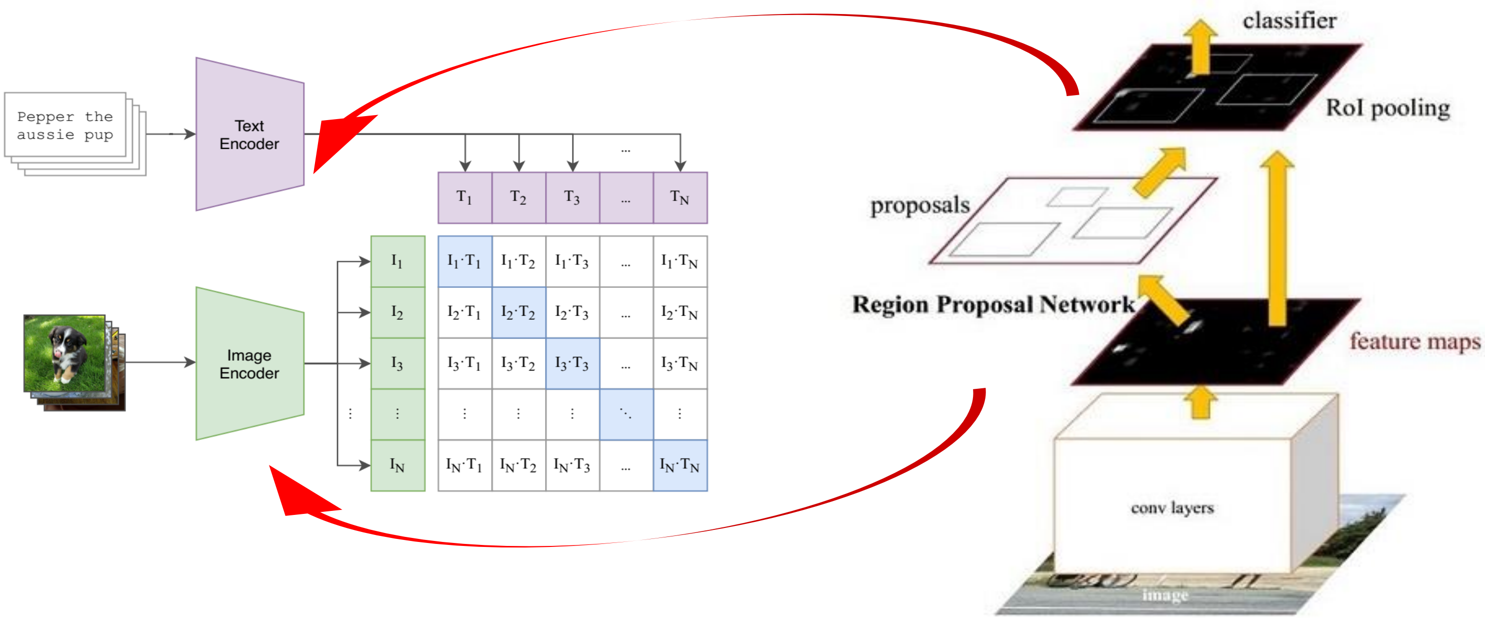


图2：CLIP模型的推理方式，可用于开放类

[1]. Radford et al. Learning transferable visual models from natural language supervision. ICML2021

研究背景二：开放类目标检测

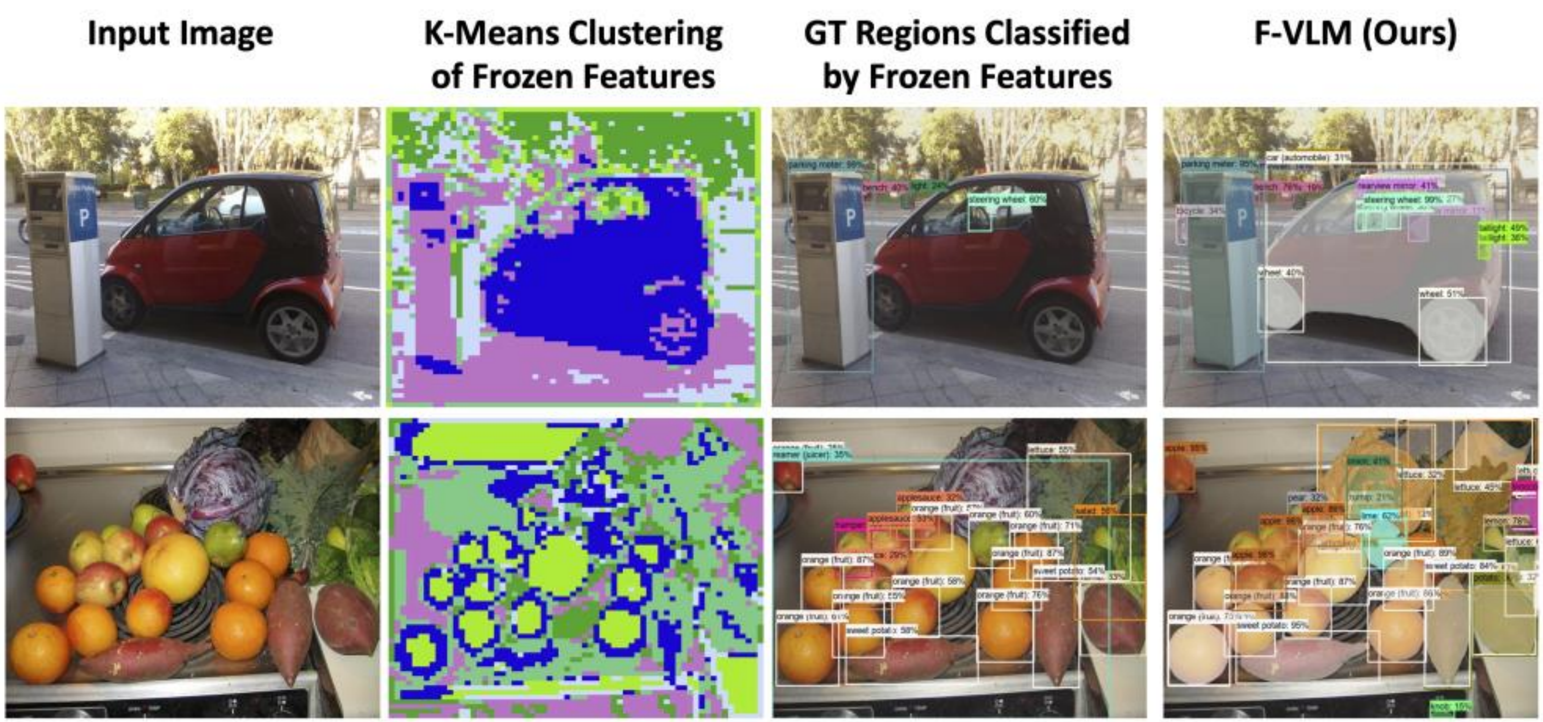
- 特点
  - 从视觉语言模型蒸馏知识
  - 结合检测模型，可以做open-vocabulary 和 zero-shot
- 难点：
  - 开放词汇需要对视觉和语言嵌入的高度理解
  - 强调泛化性就需要更大规模数据和模型



目录

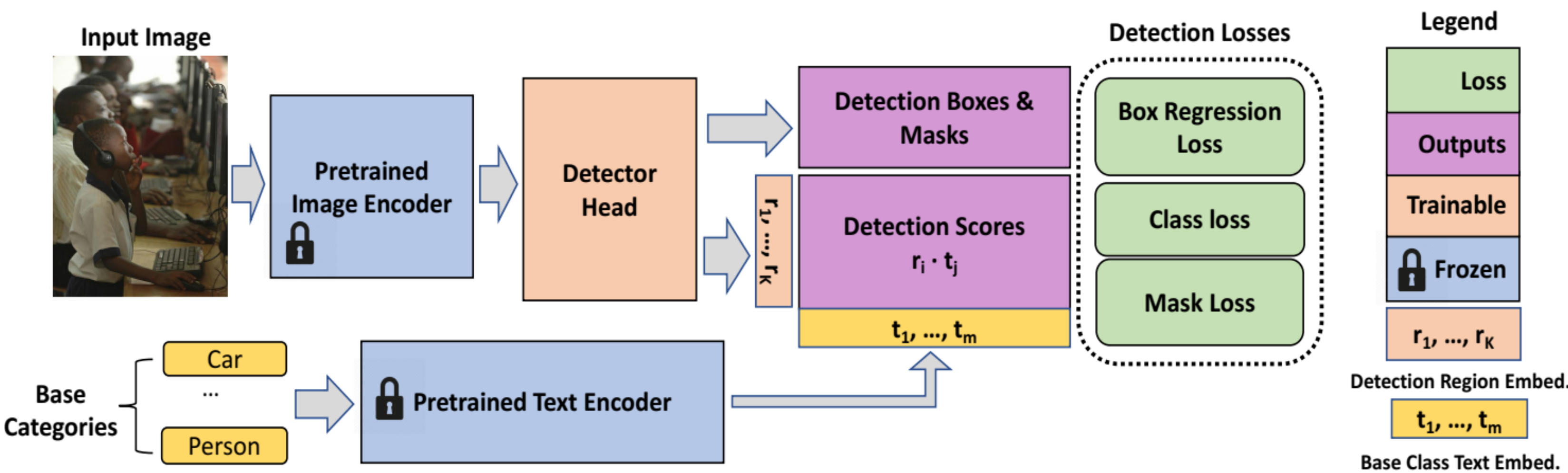
研究方法

- Motivation
  - 原始的视觉语言模型的语义提取能力和局部性很好（CLIP）
  - 冻结encoder可以减少训练和计算的消耗



研究方法

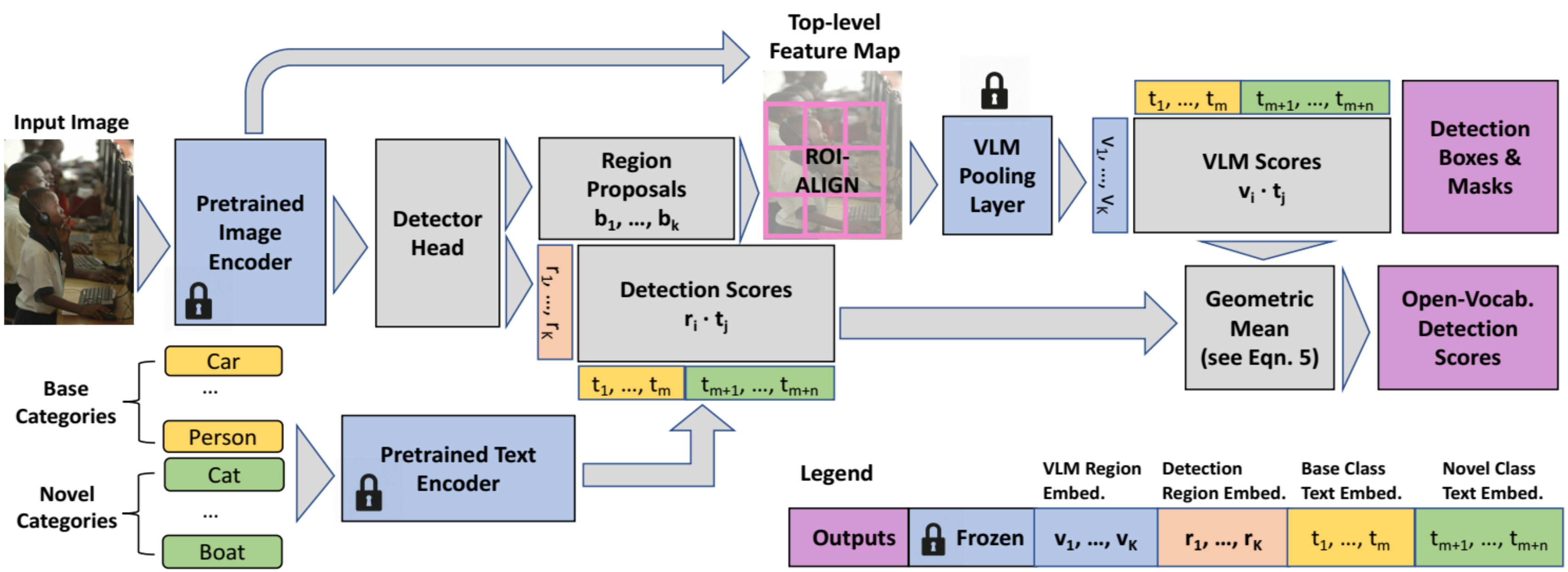
- 训练流程
  - 借鉴了Mask-RCNN的head设计方案
  - 检测得分定义为 $\mathbf{z}(\mathbf{r}_b) = Softmax(\frac{1}{\tau}\left[cos(\mathbf{r}_b, \mathbf{t}_{bg}),\ cos(\mathbf{r}_b, \mathbf{t}_1),\ \cdots,\ cos(\mathbf{r}_b, \mathbf{t}_{|C_B|})\right])$



(a) **F-VLM training architecture**. At training time, F-VLM is simply a detector with the last classification layer replaced by base-category text embeddings. The detector head is the only trainable part of the system, which includes RPN (Ren et al., 2015), FPN (Lin et al., 2017), and Mask R-CNN heads (He et al., 2017).

研究方法

- 测试流程
  - 测试过程中使用新类的语言嵌入定义视觉语言得分
  - 视觉语言得分 $\mathbf{w}(\mathbf{v}_b) = Softmax(\frac{1}{T}\left[cos(\mathbf{v}_b, \mathbf{t}_{bg}),\ cos(\mathbf{v}_b, \mathbf{t}_1),\ \cdots,\ cos(\mathbf{v}_b, \mathbf{t}_{|C_{B \cup N}|})\right])$

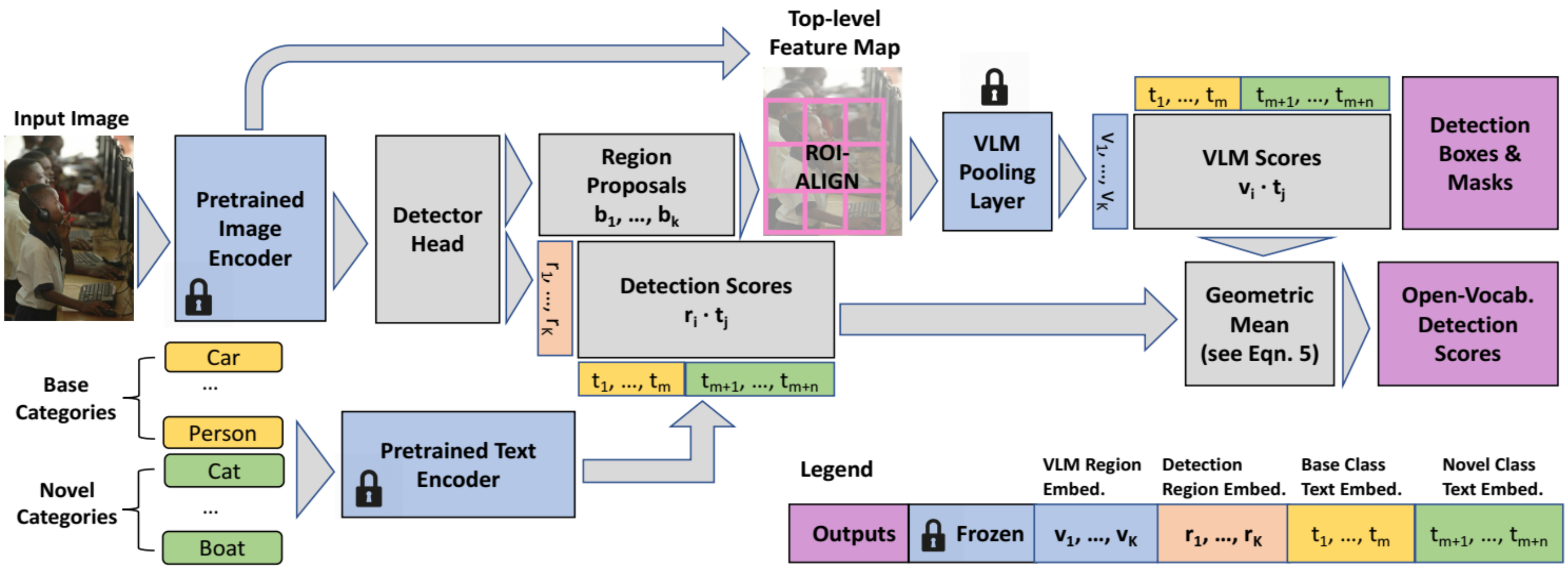


(b) **F-VLM inference architecture**. At test time, F-VLM uses the region proposals to crop out the top-level features of VLM backbone and compute the VLM score per region. The trained detector head provides the detection boxes and masks, while the classification scores are a combination of detection and VLM scores.

研究方法

- 测试流程
  - 结合检测得分和视觉语言得分，采取几何平均方案
  - 融合策略

$$s(\mathbf{r}_b)_i = \begin{cases} z(\mathbf{r}_b)_i^{(1-\alpha)} \cdot w(\mathbf{v}_b)_i^{\alpha} & \text{if } i \in C_B \\ z(\mathbf{r}_b)_i^{(1-\beta)} \cdot w(\mathbf{v}_b)_i^{\beta} & \text{if } i \in C_N \end{cases}$$



(b) **F-VLM inference architecture**. At test time, F-VLM uses the region proposals to crop out the top-level features of VLM backbone and compute the VLM score per region. The trained detector head provides the detection boxes and masks, while the classification scores are a combination of detection and VLM scores.

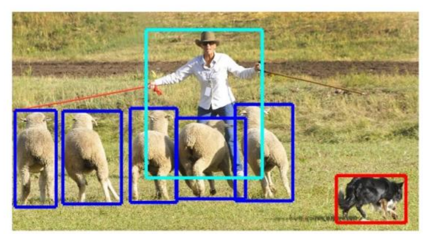
目录

实验效果

- LVIS v1 [1]
  - 866个基类（frequent & common），337个新类（rare）

- COCO [2]
  - 48个基类，17个新类，移除不包含在WordNet的15类

(a) Image classification



(b) Object localization

[1]. Gupta et al. Lvis: A dataset for large vocabulary instance segmentation. CVPR2019
[2]. Lin et al. Microsoft COCO: Common Objects in Context. ECCV2014

实验效果

LVIS数据集

Backbone (# Params)	Pretrained CLIP	Method	Distill	Trainable Backbone	AP_r	AP
R50 Comparison:						
R50	ViT-B/32	ViLD (Gu et al., 2022)	✓	✓	16.1	22.5
R50	ViT-B/32	ViLD-Ens. (Gu et al., 2022)	✓	✓	16.6	25.5
R50	ViT-B/32	DetPro (Du et al., 2022)‡	✓	✓	19.8	25.9
R50	ViT-B/32	Detic-ViLD (Zhou et al., 2022c)*	✗	✓	17.8	26.8
R50	R50	RegionCLIP (Zhong et al., 2022)†	✓	✓	17.1	28.2
R50	R50	F-VLM (Ours)	✗	✗	18.6	24.2
System-level Comparison:						
R152 (60M)	ViT-B/32	ViLD (Gu et al., 2022)	✓	✓	18.7	23.6
R152 (60M)	ViT-B/32	ViLD-Ens. (Gu et al., 2022)	✓	✓	18.7	26.0
EN-B7 (67M)	ViT-L/14	ViLD-Ens. (Gu et al., 2022)	✓	✓	21.7	29.6
EN-B7 (67M)	EN-B7*	ViLD-Ens. (Gu et al., 2022)	✓	✓	26.3	29.3
R50 (26M)	ViT-B/32	DetPro-Cascade (Du et al., 2022)‡	✓	✓	20.0	27.0
R50 (26M)	ViT-B/32	Detic-CN2 (Zhou et al., 2022c)*	✗	✓	24.6	32.4
R50x4 (87M)	R50x4	RegionCLIP (Zhong et al., 2022)†	✓	✓	22.0	32.3
ViT-L/14 (303M)	ViT-L/14	OWL-ViT (Minderer et al., 2022)	✗	✓	25.6	34.7
R50x4 (87M)	R50x4	F-VLM (Ours)	✗	✗	26.3	28.5
R50x16 (167M)	R50x16	F-VLM (Ours)	✗	✗	30.4	32.1
R50x64 (420M)	R50x64	F-VLM (Ours)	✗	✗	**32.8**	34.9

实验效果

- COCO数据集

Method	Training source	Novel AP	AP
WSDDN (Bilen & Vedaldi, 2016)	image-level labels in $C_B \cup C_N$	19.7	19.6
Cap2Det (Ye et al., 2019)		20.3	20.1
ZSD (Bansal et al., 2018)	instance-level labels in C_B	0.31	24.9
DELO (Zhu et al., 2020)		3.41	13.0
PL (Rahman et al., 2020)		4.12	27.9
OVR-CNN (Zareian et al., 2021)	image captions in $C_B \cup C_N$ instance-level labels in C_B	22.8	39.9
CLIP-RPN (Gu et al., 2022)	CLIP image-text pairs instance-level labels in C_B	26.3	27.8
ViLD (Gu et al., 2022)		27.6	51.3
Detic* (Zhou et al., 2022c)		27.8	45.0
RegionCLIP‡ (Zhong et al., 2022)		**31.4**	50.4
RegionCLIP† (Zhong et al., 2022)		26.8	47.5
RegionCLIP⋆ (Zhong et al., 2022)		14.2	42.7
F-VLM (Ours)		28.0	39.6

实验效果

- 训练需求

Table 3: **Training Resource Benchmark.** We report LVIS mask AP_r to show the performance vs training cost trade-off. F-VLM can outperform ViLD (Gu et al., 2022) with 226× less compute.

Method	Mask AP_r	#Iters	Epochs	Training Cost (Per-Core-Hour)	Training Cost Savings
ViLD-EN-B7 (Gu et al., 2022)	26.3	180k	460	8000	1×
F-VLM (Ours)	32.8	46.1k	118	565	14×
F-VLM (Ours)	31.0	5.76k	14.7	71	113×
F-VLM (Ours)	27.7	**2.88k**	**7.4**	**35**	**226×**

实验效果

- 迁移效果

Table 4: **Transfer detection of F-VLM**. We evaluate LVIS-trained F-VLM on COCO and Objects365 without finetuning. F-VLM demonstrates strong scaling property with a gain of +7.3/+5.8 AP on COCO/Objects365 by increasing backbone capacity.

Method	COCO			Objects365		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised (Gu et al., 2022)	46.5	67.6	50.9	25.6	38.6	28.0
ViLD-R50 (Gu et al., 2022)	36.6	55.6	39.8	11.8	18.2	12.6
DetPro-R50 (Du et al., 2022)	34.9	53.8	37.4	12.1	18.8	12.9
F-VLM-R50 (Ours)	32.5	53.1	34.6	11.9	19.2	12.6
F-VLM-R50x4 (Ours)	36.0	57.5	38.7	14.2	22.6	15.2
F-VLM-R50x16 (Ours)	37.9	59.6	41.2	16.2	25.3	17.5
F-VLM-R50x64 (Ours)	**39.8**	**61.6**	**43.8**	**17.7**	**27.4**	**19.1**

实验效果

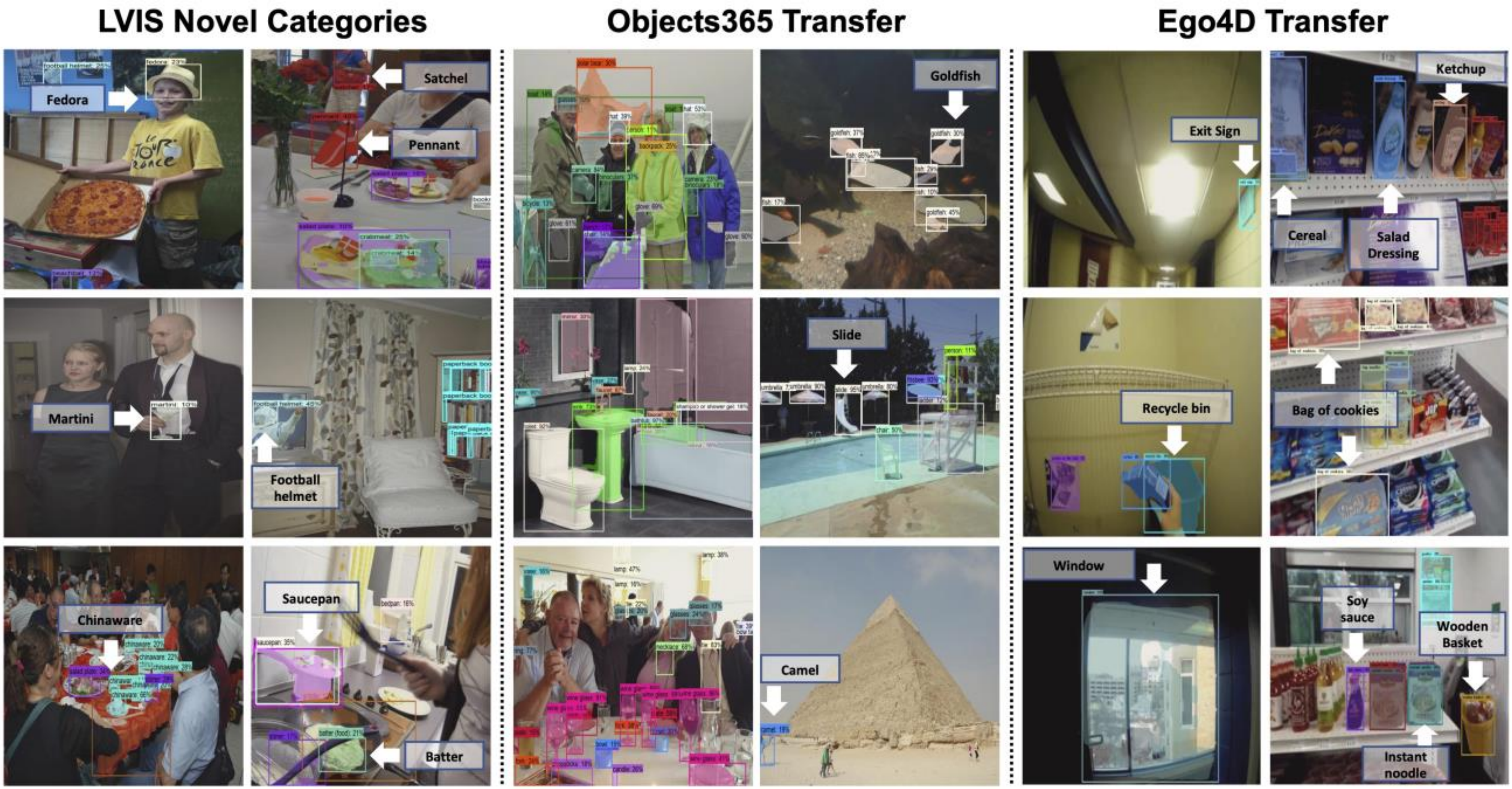


图1：迁移到其他数据集上的可视化实验。

目录

总结

总结

- 本文和很多OVOD方法落脚点相似，MaskRCNN+CLIP：
  - ViLD，DetPro等方法通过知识蒸馏
  - GLIP，OVD-ViT走大规模预训练+finetuen策略
- 本文特点在于：确立了固定backbone的思路保持模型泛化能力，因为finetune本身对域外数据不友好（如表5）
- 借助大模型能力对于零样本、长尾任务有降维打击效果

Table 5: **Finetuning vs frozen backbone.** Finetuning does not benefit the novel categories (AP_r) but improves the base categories (AP_c, AP_f).

Backbone LR	AP_r	AP_c	AP_f	AP
1e-3	18.1	25.7	30.2	26.2
1e-4	18.1	24.9	28.8	25.3
0.0	18.6 (+0.5)	24.0	26.9	24.2