北京大学
PEKING UNIVERSITY

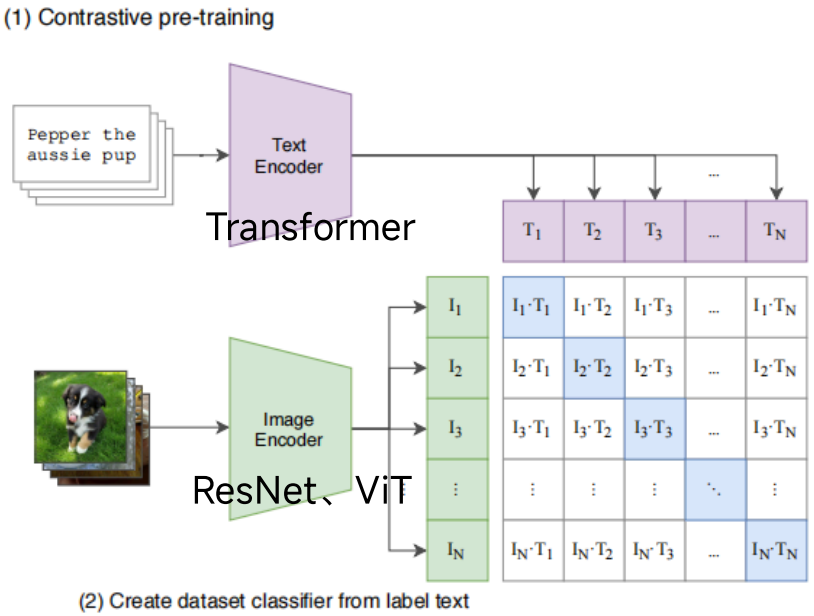
Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation

Google research
ICLR 2022

彭天天

2024/11/29

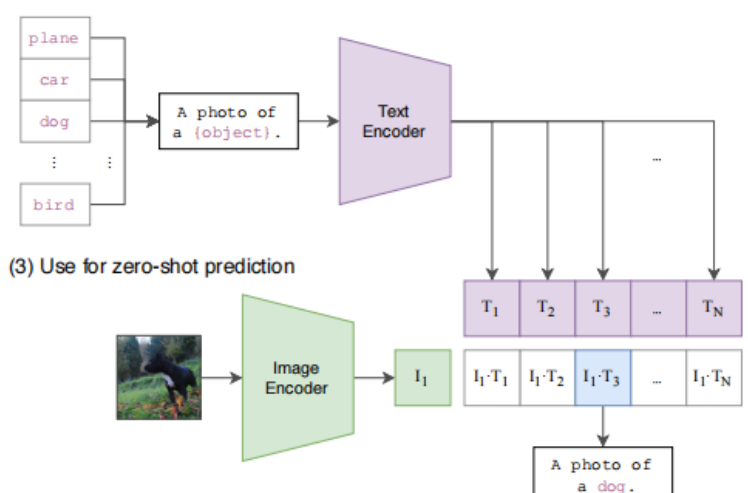
训练



输入：图片-文本对。
计算文本和图像嵌入的相似度，计算交叉熵loss

数据集：WebImageText，包含4亿个图像-文本对

推理



可直接进行zero-shot的图像分类

根据label构建描述文本，如“ A photo of {label} ”，计算图片与文本嵌入的相似度（logits），送入softmax得到每个类别的预测概率

Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

经典的目标检测两阶段结构：Fast R-CNN

先使用backbone提取图像特征features，由RPN提出proposals
再将特征送入RoIHeads输出分类结果和边界框

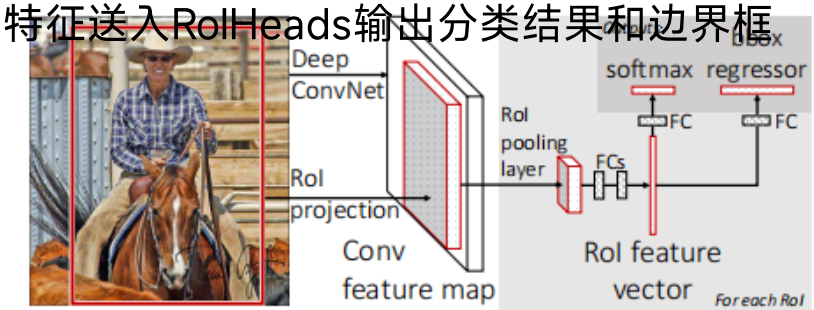
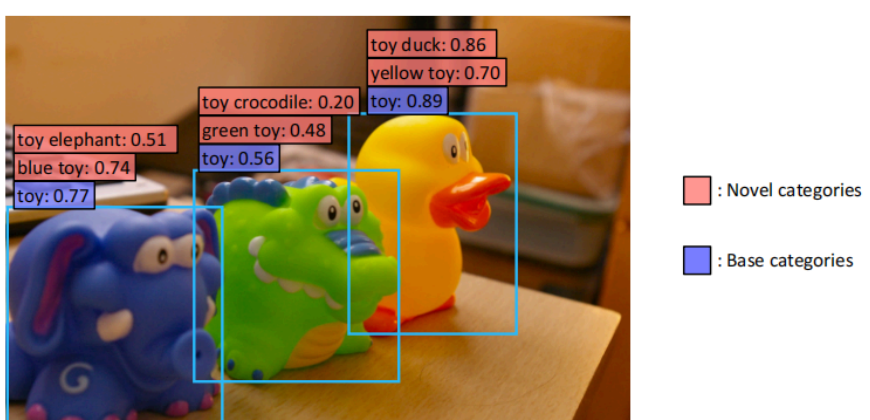


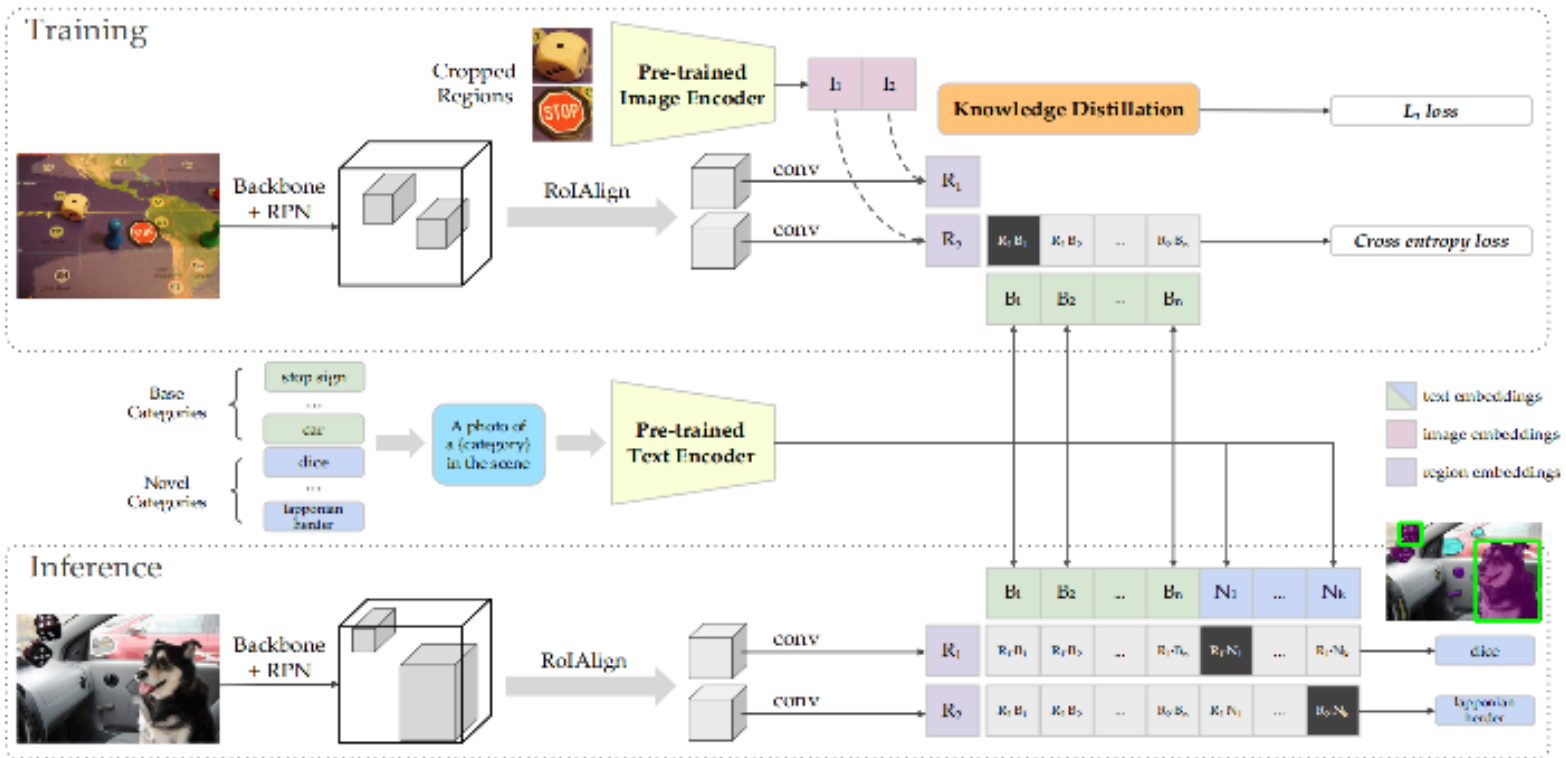
Figure 1. Fast R-CNN architecture. An input image and multiple regions of interest (RoIs) are input into a fully convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs). The network has two output vectors per RoI: softmax probabilities and per-class bounding-box regression offsets. The architecture is trained end-to-end with a multi-task loss.

Figures from: Fast R-CNN

改进，通过知识蒸馏从 CLIP 中学习，从而在推理时能够检测到任意的新的物体类别

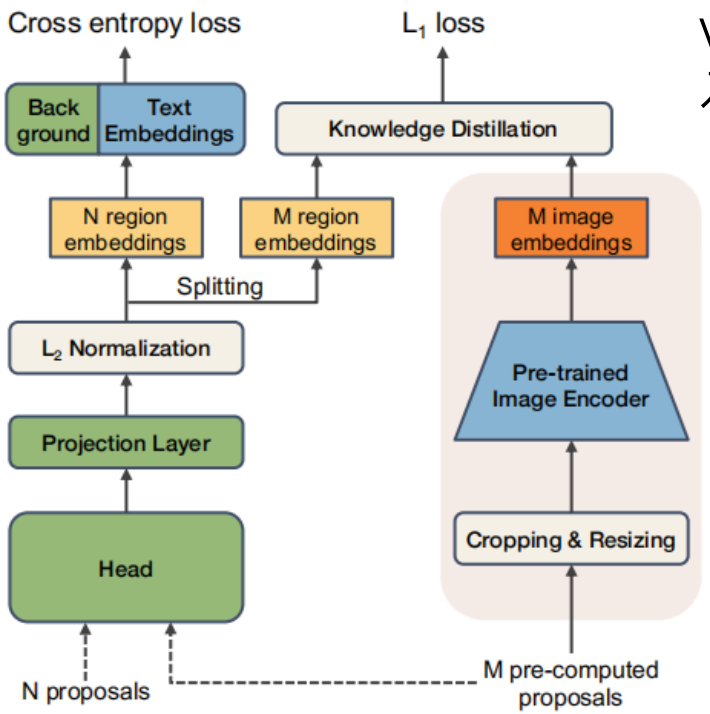
1.将proposals区域的RoI feature与CLIP的图像嵌入对齐。
2.使用文本嵌入和区域嵌入对齐来替换分类头





只使用base类进行训练

推理时可根据开放词汇表进行目标检测



(d) ViLD

1.ViLD-image

在base类上训练一个RPN，并选择其建议的前M个proposals $\tilde{r} \in \tilde{P}$

V代表CLIP，$\mathcal{V}(\text{crop}(I, \tilde{r}))$ 表示图像I经过proposals裁剪后送入CLIP得到的图像嵌入(offline)

$$\mathcal{V}(\text{crop}(I, \tilde{r}_{\{1\times,1.5\times\}})) = \frac{\mathbf{v}}{\|\mathbf{v}\|}, \text{ where } \mathbf{v} = \mathcal{V}(\text{crop}(I, \tilde{r}_{1\times})) + \mathcal{V}(\text{crop}(I, \tilde{r}_{1.5\times})). \quad (1)$$

$$\mathcal{L}_{\text{ViLD-image}} = \frac{1}{M} \sum_{\tilde{r}\in\tilde{P}} \|\mathcal{V}(\text{crop}(I, \tilde{r}_{\{1\times,1.5\times\}})) - \mathcal{R}(\phi(I), \tilde{r})\|_1. \quad (3)$$

2.ViLD-text

计算区域嵌入与类别嵌入的相似性，计算分类结果与交叉熵loss（online）

$$\mathbf{e}_r = \mathcal{R}(\phi(I), r)$$

$$\mathbf{z}(r) = \left[sim(\mathbf{e}_r, \mathbf{e}_{bg}),\ sim(\mathbf{e}_r, \mathbf{t}_1),\ \cdots,\ sim(\mathbf{e}_r, \mathbf{t}_{|C_B|})\right]$$

$$\mathcal{L}_{\text{ViLD-text}} = \frac{1}{N} \sum_{r\in P} \mathcal{L}_{\text{CE}}\left(softmax(\mathbf{z}(r)/\tau), y_r\right),$$

$$\mathcal{L}_{\text{ViLD}} = \mathcal{L}_{\text{ViLD-text}} + w \cdot \mathcal{L}_{\text{ViLD-image}},$$

Table 1: **Training with only base categories achieves comparable average recall (AR) for novel categories on LVIS.** We compare RPN trained with base only *vs.* base+novel categories and report the bounding box AR.

Supervision	AR_r@100	AR_r@300	AR_r@1000
base	39.3	48.3	55.6
base + novel	41.1	50.9	57.0

Table 3: **Performance of ViLD and its variants. ViLD outperforms the supervised counterpart on novel categories.** Using ALIGN as the teacher model achieves the best performance without bells and whistles. All results are mask AP. We average over 3 runs for R50 experiments. †: methods with R-CNN style; runtime is 630× of Mask R-CNN style. ‡: for reference, fully-supervised learning with additional tricks.

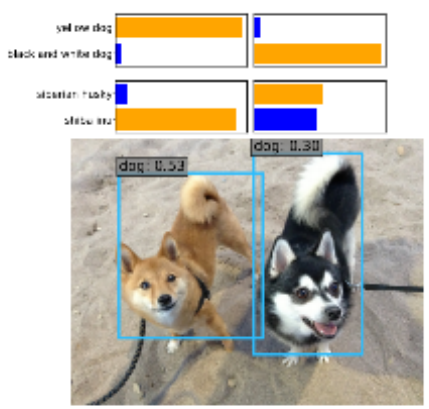
Backbone	Method	AP_r	AP_c	AP_f	AP
ResNet-50+ViT-B/32	CLIP on cropped regions†	18.9	18.8	16.0	17.7
	ViLD-text+CLIP†	**22.6**	24.8	29.2	26.1
ResNet-50	Supervised-RFS (base+novel)	12.3	24.3	32.4	25.4
	GloVe baseline	3.0	20.1	30.4	21.2
	ViLD-text	10.1	23.9	32.5	24.9
	ViLD-image	11.2	11.3	11.1	11.2
	ViLD (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
	ALIGN on cropped regions	**39.6**	32.6	26.3	31.4

- **RPN的影响**
- RPN在base类上训练和在base+novel上训练，对于rare类（即novel类）的候选框召回率影响不大
- 更好的RPN也许能保留更多的unseen目标

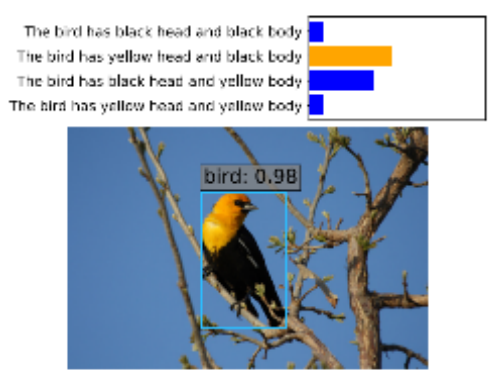
- **与其他方法的比较**
- 同backbone前提下，ViLD比监督学习的效果要好
- backbone越大越强
- 教师越强，蒸馏效果越好（CLIP到ALIGN）

Table 5: **Generalization ability of ViLD.** We evaluate the LVIS-trained model with ResNet-50 backbone on PASCAL VOC 2007 test set, COCO validation set, and Objects365 v1 validation set. Simply replacing the text embeddings, our approaches are able to transfer to various detection datasets. The supervised baselines of COCO and Objects365 are trained from scratch. †: the supervised baseline of PASCAL VOC is initialized with an ImageNet-pretrained checkpoint. All results are box APs.

Method	PASCAL VOC†		COCO			Objects365		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
ViLD-text	40.5	31.6	28.8	43.4	31.4	10.4	15.8	11.1
ViLD	72.2	56.7	36.6	55.6	39.8	11.8	18.2	12.6
Finetuning	78.9	60.3	39.1	59.8	42.4	15.2	23.9	16.2
Supervised	78.5	49.0	46.5	67.6	50.9	25.6	38.6	28.0



(a) Fine-grained breeds and colors.

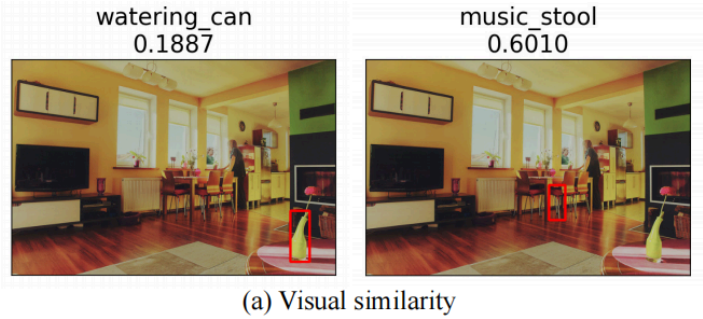


(b) Colors of body parts.

Figure 5: **On-the-fly interactive object detection.** One application of ViLD is using on-the-fly arbitrary texts to further recognize more details of the detected objects, *e.g.*, fine-grained categories and color attributes.

- **数据集迁移（在LVIS上训练）**
- ViLD泛化能力较强

- **交互式目标检测**
- 可以通过更改文本可获取目标的更多信息（类别、颜色等）
- 对颜色敏感，对动作等抽象的概念不敏感



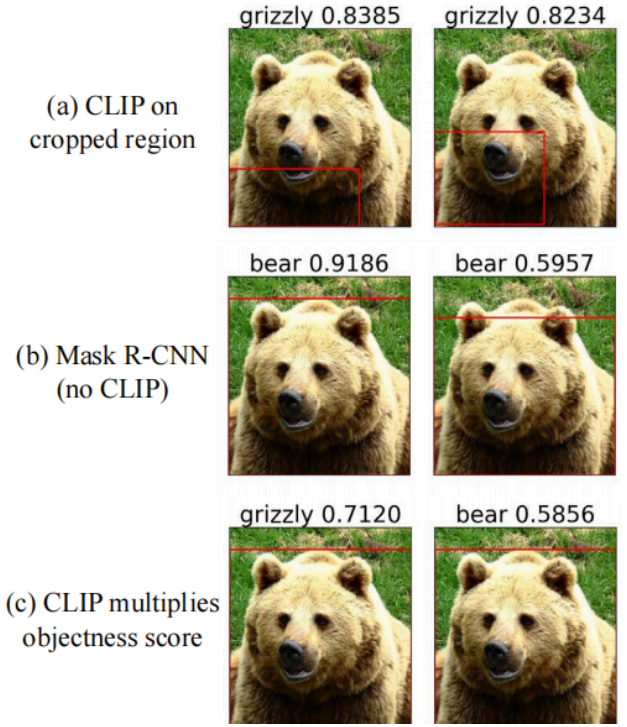
(a) Visual similarity

(b) Aspect ratio

(c) Other objects in the bounding box

- **一些表现不佳的例子**
- a）视觉上相似的物体难以区分
- b）由于resize产生的误解——region可能是长条形的，resize成224×224后可能与其他正方形的物体相似
- c）由clip本身带来的偏差——倾向于表

- **只用CLIP定位，质量差**
- 与文本相似度最高的框不一定是最好的框，综合RPN分数得出最好的框



除此之外还比较了不同数据集、loss、超参值、prompt的影响

CoDet: Co-Occurrence Guided Region-Word Alignment for Open-Vocabulary Object Detection

The University of Hong Kong ByteDance Inc

NeurIPS 2023

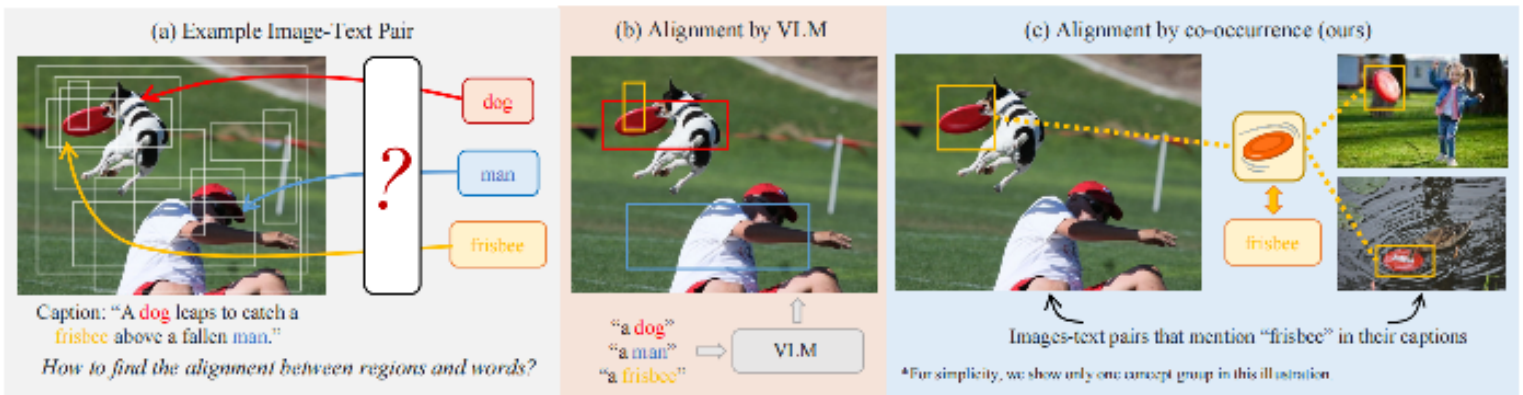
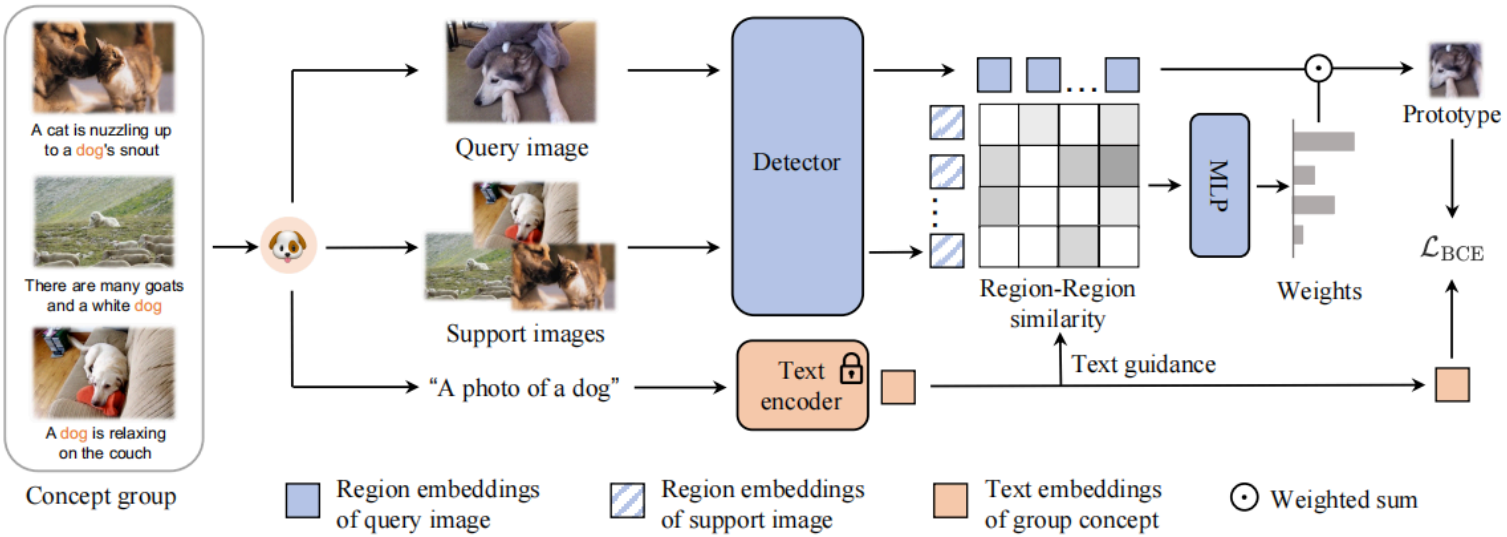


Figure 1: **Illustration of different region-text alignment paradigms.** (a): example image-text pair, and region proposals generated by a pre-trained region proposal network; (b): a pre-trained VLM (*e.g.*, CLIP [35]) is used to retrieve the box with the highest region-word similarity concerning the query text, which yet exhibits poor localization quality; (c) our method overcomes the reliance on VLMs by exploring visual clues, *i.e.*, object co-occurrence, within a group of image-text pairs containing the same concept (*e.g.*, frisbee). *Best viewed in color.*

- b）在众多proposals中，与文本嵌入相似度最高的框不一定是最好的框
- c）通过co-occurrence（共现）对齐后能更准确地掌握concept-region关系



1.从数据集中<I, T>提取所有的concept words，并根据concept words将数据集分成若干个concept Group（一个图像可以属于多个组），每次训练从一个组中抽取m+1张图像进行训练

模型的目标就是将这些组中重复出现的目标与concept对应上

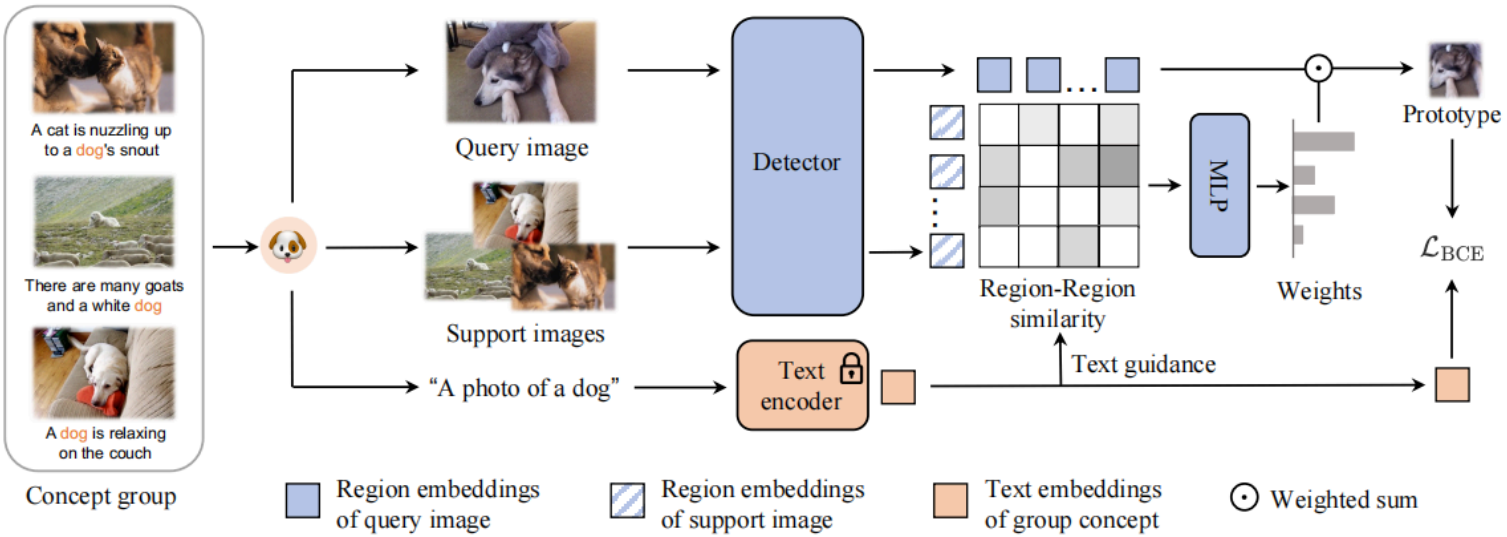
2.m+1张图像迭代地选取其中一张作为query image，剩下m张作为support images，送入Detector生成region级特征，计算一个 $S^{n\times mn}$ 相似度矩阵 $\mathbf{f}_p = \sum_{i=1}^{N} \mathbf{p}_i \cdot \mathbf{f}_i, \quad \text{where } \mathbf{p} = \underset{N}{\text{softmax}}\left(\Phi\left(\mathbf{S}\right)\right)$

fp 代表 $\mathcal{L}_{\text{region-word}} = \mathcal{L}_{\text{BCE}}(\mathbf{W}\mathbf{f}_p, c)$ 可能包含目标concept的region的综合特征

3.文本引导相似度矩阵S生成

$$s_{ij} = \bar{\mathbf{w}}_c^{\top} \cdot \left(\frac{\mathbf{f}_i}{\|\mathbf{f}_i\|} \circ \frac{\mathbf{f}_j}{\|\mathbf{f}_j\|} \right)$$

避免多个共现对象的干扰（比如所有图片既有猫又有狗）



训练过程：

detection data和image-text pairs轮流输入(如2:2×4)

前者用于训练模型的目标检测能力

后者用于视觉与文本的对齐。

$L_{\text{image-text}}$是将整个图像作为image，类似clip

$$\mathcal{L}(I) = \begin{cases} \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}}, & \text{if } I \in \mathcal{D}^{\text{det}} \\ \mathcal{L}_{\text{region-word}} + \mathcal{L}_{\text{image-text}}, & \text{if } I \in \mathcal{D}^{\text{cap}} \end{cases}$$

Table 1: **Comparison with state-of-the-art open-vocabulary object detection methods on OV-LVIS.** Caption supervision means the method learns vision-language alignment from image-text pairs, while CLIP supervision indicates transferring knowledge from pre-trained CLIP. The column 'Strict' indicates whether the method follows a strict open-vocabulary setting.

Method	Backbone	Supervision	Strict	AP^m_{novel}	AP^m_c	AP^m_f	AP^m_{all}
ViLD [18]	RN50-FPN	CLIP	✓	16.6	24.6	30.3	25.5
RegionCLIP [59]	RN50-C4	Caption	✓	17.1	27.4	34.0	28.2
DetPro [12]	RN50-FPN	CLIP	✓	19.8	25.6	28.9	25.9
OV-DETR [56]	RN50-C4	Caption	✗	17.4	25.0	32.5	26.6
PromptDet [14]	RN50-FPN	Caption	✗	19.0	18.5	25.8	21.4
Detic [61]	RN50	Caption	✗	19.5	-	-	30.9
F-VLM [26]	RN50-FPN	CLIP	✓	18.6	-	-	24.2
VLDet [28]	RN50	Caption	✓	21.7	29.8	34.3	30.1
BARON [50]	RN50-FPN	CLIP	✓	22.6	27.6	29.8	27.6
CoDet (Ours)	RN50	Caption	✓	**23.4**	30.0	34.6	30.7
RegionCLIP [59]	R50x4 (87M)	Caption	✓	22.0	32.1	36.9	32.3
Detic [61]	SwinB (88M)	Caption	✗	23.9	40.2	42.8	38.4
F-VLM [26]	R50x4 (87M)	CLIP	✓	26.3	-	-	28.5
VLDet [28]	SwinB (88M)	Caption	✓	26.3	39.4	41.9	38.1
CoDet (Ours)	SwinB (88M)	Caption	✓	**29.4**	39.5	43.0	39.2
F-VLM [26]	R50x64 (420M)	CLIP	✓	32.8	-	-	34.9
CoDet (Ours)	EVA02-L (304M)	Caption	✓	**37.0**	46.3	46.3	44.7

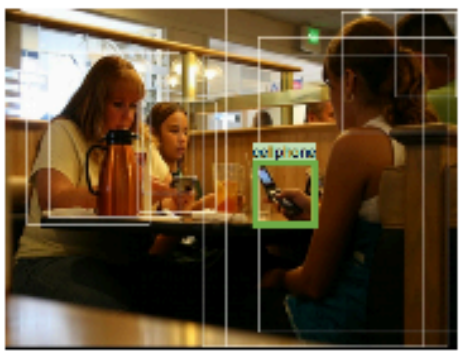
Table 2: **Comparison with state-of-the-art methods on OV-COCO.** †: implemented with Deformable DETR [64].

Method	AP^{novel}_{50}	AP^{base}_{50}	AP^{all}_{50}
OVR-CNN [57]	22.8	46.0	39.9
ViLD [18]	27.6	59.5	51.3
RegionCLIP [59]	26.8	54.8	47.5
Detic [61]	27.8	47.1	42.0
OV-DETR [56]†	29.4	61.0	52.7
PB-OVD [15]	29.1	44.4	40.4
VLDet	**32.0**	50.6	45.8
CoDet (Ours)	<u>30.6</u>	52.3	46.6

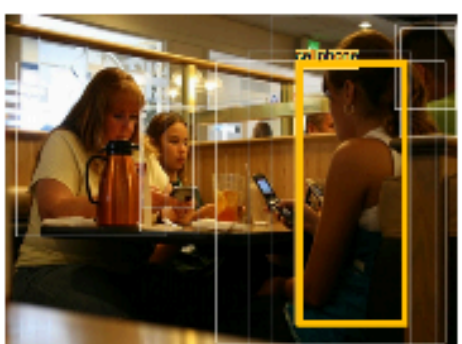
Table 3: **Cross-datasets transfer detection from OV-LVIS to COCO and Objects365.** †: Detection-specialized pre-training with SoCo [48].

Method	COCO			Objects365		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised [18]	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [18]	36.6	55.6	39.8	11.8	18.2	12.6
DetPro [12]†	34.9	53.8	37.4	12.1	18.8	12.9
F-VLM [26]	32.5	53.1	34.6	11.9	19.2	12.6
BARON [50]	36.2	55.7	39.1	13.6	**21.0**	14.5
CoDet (Ours)	**39.1**	**57.0**	**42.3**	**14.2**	20.5	**15.3**

region-region

region-text

region-region对齐，bbox效果优于region-text对齐

- 文本引导相似度矩阵生成

(a) **Text guidance**

Text guide	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
✗	26.6	52.4	45.7
✓	30.6	52.3	46.6

w/o text guidance　　w/ text guidance “pigeon”

Figure 5: There can be more than one co-occurring concept among sampled images. Text guidance helps filter out the distracting concept (chair legs) and focus on the concept of interest (pigeons).

北京大学
PEKING UNIVERSITY

Open-Vocabulary SAM: Segment and Recognize Twenty-thousand Classes Interactively

S-Lab, Nanyang Technological University Shanghai AI Laboratory

ECCV 2024

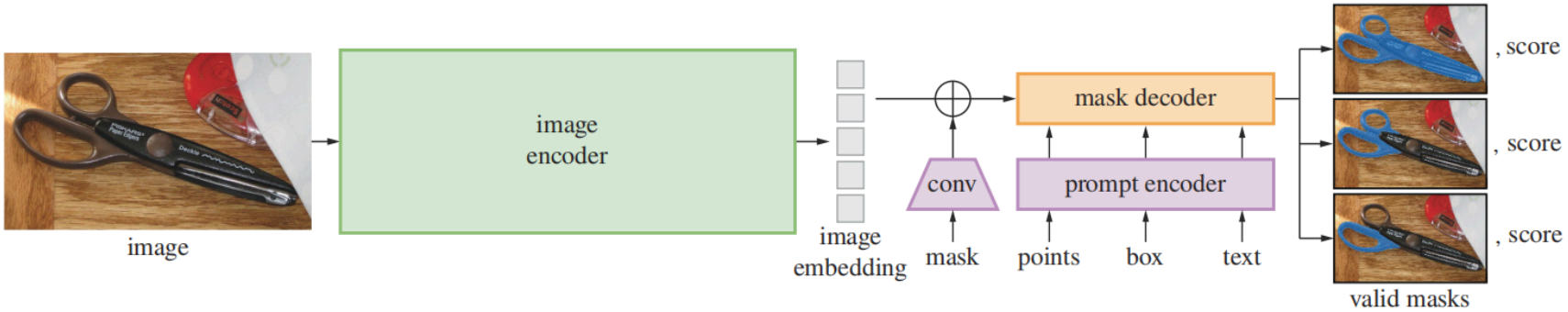


Image encoder:
MAE pre-trained Vision Transformer (ViT)

Prompt encoder:
points, box
text (CLIP)
mask (conv)

Mask decoder:
based on Transformer decoder block
output masks and scores

Zero-Shot Text-to-Mask:

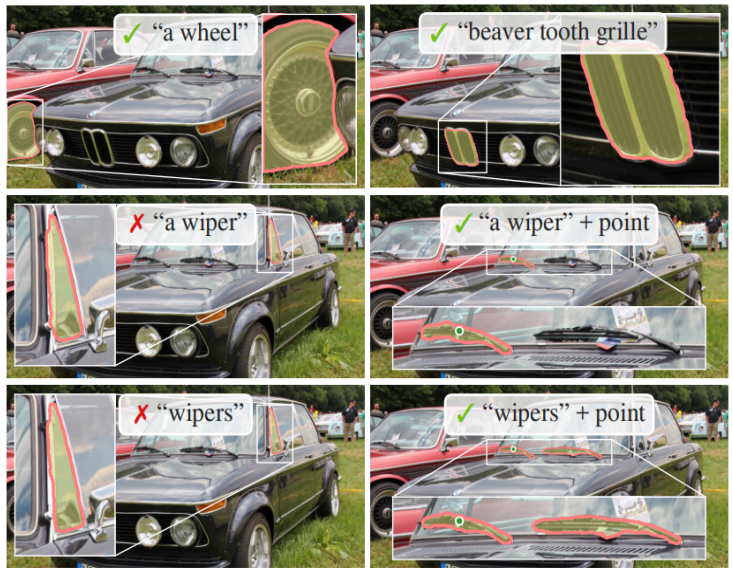


Figure 8: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Segment Anything
by Meta AI, ICCV 2023

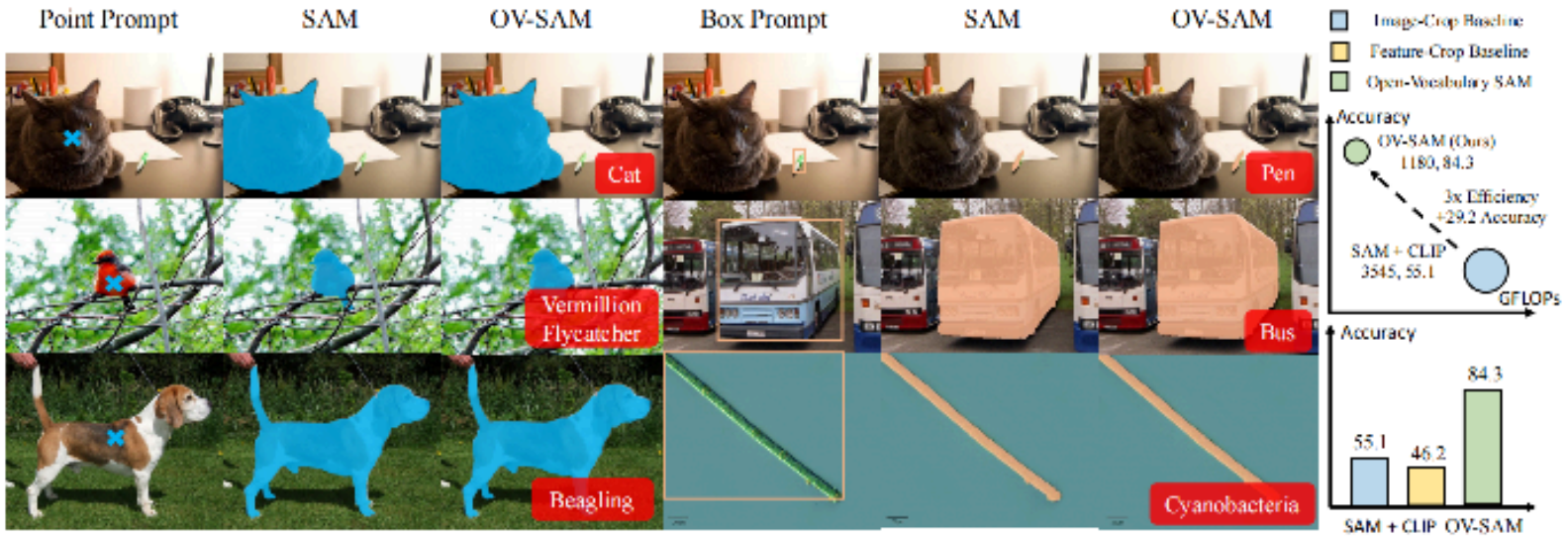


Fig. 1: Open-Vocabulary SAM not only can segment anything with prompts just like SAM but also has the capability of recognition in the real world, like CLIP. With drastically lower computational cost, Open-Vocabulary SAM has a higher recognition performance than directly combining SAM and CLIP with image or feature cropping (measured on the COCO open vocabulary benchmark).

给定图像和prompt
SAM: 输出mask
OV-SAM: 输出mask和label

表现优于SAM与CLIP的直接组合

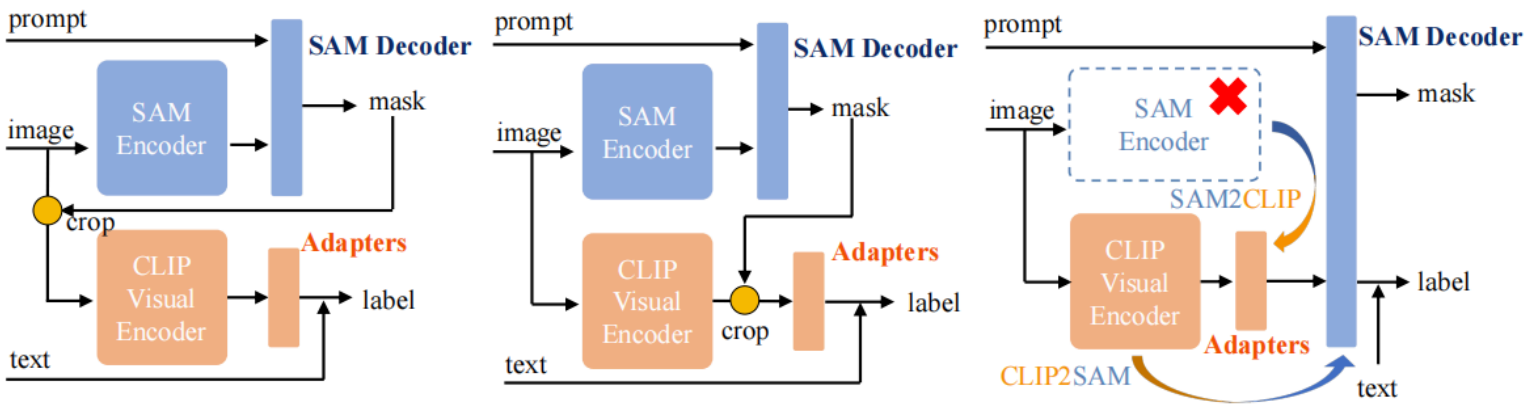
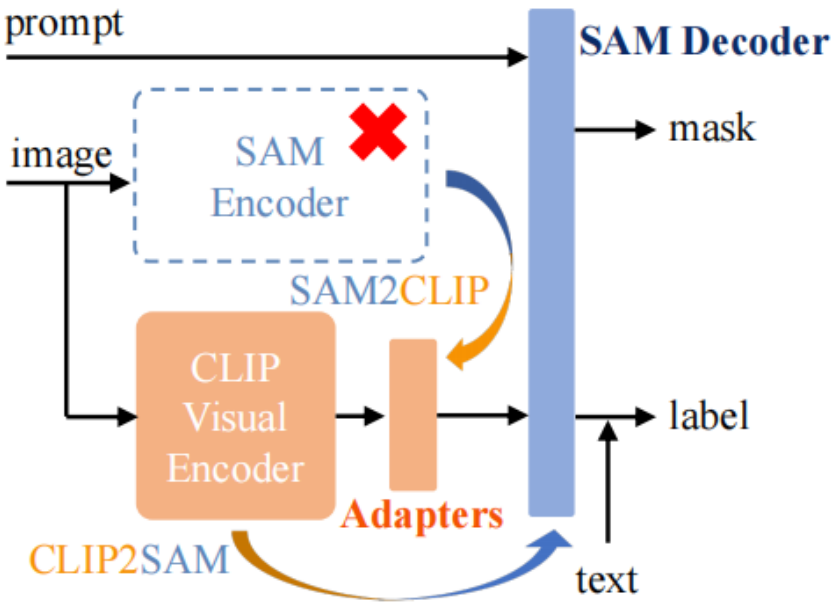


Fig. 2: Comparison of two simple SAM-CLIP combination baselines (a) and (b), and our proposed single encoder architecture (c). The adapters for (a) and (b) are optional and can be replaced with various designs (please refer to Sec. 4.1 for details). Note that, in our method, the SAM encoder will be discarded during inference.

两种常见的SAM-CLI简单组合与SAM对比

SAM2CLIP：SAM Encoder作为教师训练adapters，将CLIP提取的视觉特征与SAM对齐；

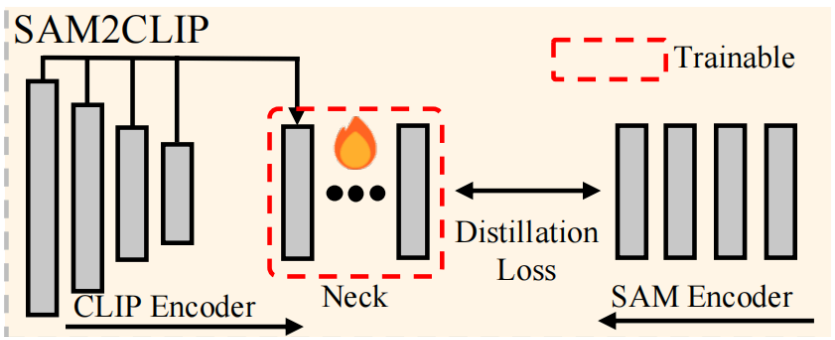
CLIP2SAM：对齐后的特征使用SAM的解码器来进行分割



(c) Open-Vocabulary SAM

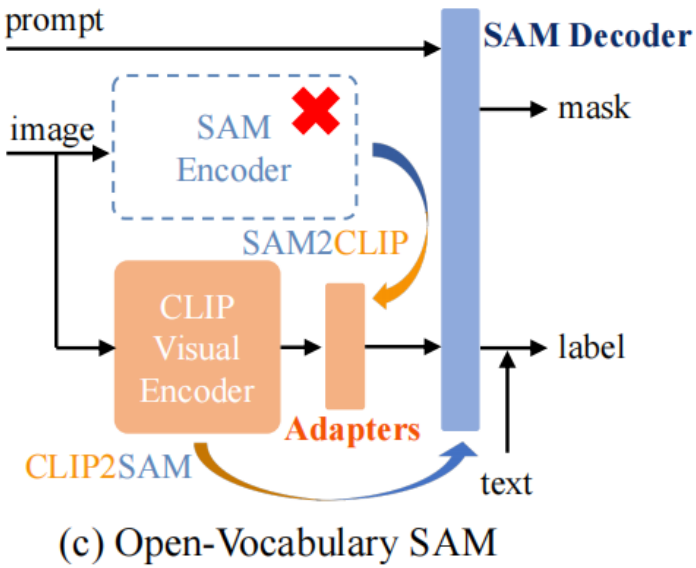
- 第一阶段：特征提取

提取CLIP文本嵌入和SAM特征并保存



- 第二阶段：SAM2CLIP

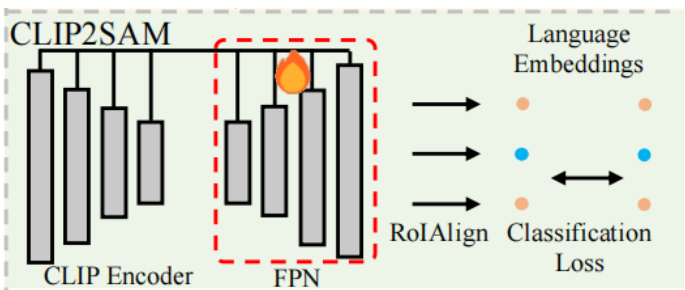
$$L_{distill} = \mathrm{MSE}(F_{sam}, A_{sam2clip}(\mathrm{Fusion}(E_I^i))),$$



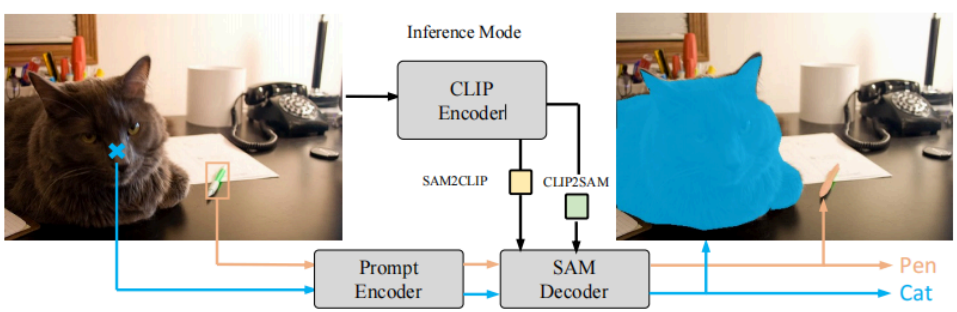
(c) Open-Vocabulary SAM

- 第三阶段：SAM Decoder、CLIP2SAM

为了解决对小目标的分类识别，引入FPN网络+conv+mlp，得到的特征一起送入SAM decoder



$$L=L_{cls}+L_{mask\_ce}+L_{mask\_dice}$$



: Parameters Trainable : Parameters Frozen

- SAM encoder is dropped during the inference.
- Use open-vocabulary classifier.
- CLIP encoder is only inference once.
- Interactive segmentation and recognition

Table 1: Comparison of combined baselines and Open-Vocabulary SAM using visual prompts. "*" indicates using mask center point as prompts, while others indicate using ground truth boxes prompts. IoU_b and IoU_n refer to the average IoU for each mask of base classes and novel classes, respectively.

Method	COCO			LVIS			FLOPs	#Param
	IoU_b	IoU_n	Acc	IoU_b	IoU_n	Acc		
Image-Crop baseline	78.1	81.4	46.2	78.3	81.6	9.6	3,748G	808M
Feature-Crop baseline	78.1	81.4	55.1	78.3	81.6	26.5	3,545G	808M
Image-Crop baseline + CoOp [89]	79.6	82.1	62.0	80.1	82.0	32.1	3,748G	808M
Feature-Crop baseline + CoOp [89]	79.6	82.1	70.9	80.1	82.0	48.2	3,545G	808M
Open-Vocabulary SAM	81.5	84.0	84.3	80.4	83.1	66.6	1,180G	304M
Image-Crop baseline*	60.7	66.7	24.5	53.0	62.3	6.2	3,748G	808M
Feature-Crop baseline*	60.7	66.7	32.1	53.0	62.3	11.0	3,545G	808M
Image-Crop baseline + CoOp [89]*	64.7	66.7	28.2	58.9	64.2	8.3	3,748G	808M
Feature-Crop baseline + CoOp [89]*	64.7	66.7	35.1	58.9	64.2	13.2	3,545G	808M
Open-Vocabulary SAM*	68.4	65.2	76.7	63.6	67.9	60.4	1,180G	304M

Table 2: Comparison of combined baselines and Open-Vocabulary SAM on prompts generated by the open vocabulary detector. For the LVIS dataset, only 'normal' and 'frequent' classes are in the training set. The labels are generated by each baseline or our method. We adopt Detic [90] as the OV-Detector to provide box prompts.

Method	COCO			LVIS				FLOPs	#Params
	AP_{base}	AP_{novel}	AP	AP_{rare}	AP_{norm}	AP_{freq}	AP		
Image-Crop baseline + CoOp [89]	26.2	31.2	27.3	19.8	18.3	16.3	17.2	3,748G	808M
Feature-Crop baseline + CoOp [89]	28.0	33.8	29.5	24.2	21.4	18.6	20.8	3,545G	808M
Open-Vocabulary SAM	31.1	36.0	32.4	24.0	21.3	22.9	22.4	1,180G	304M

- 与SAM+CLIP表现对比

性能表现接近最优，但参数量减少一倍多

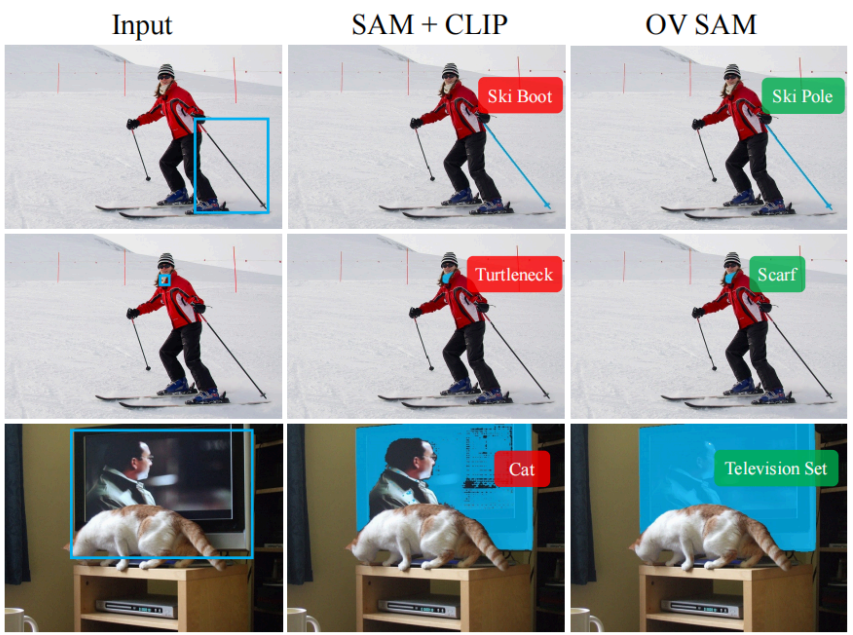


Table 4: Comparison of mask quality with various detectors on COCO dataset. We report the mask mean AP for comparison. The masks are generated by each method, while the labels are from the corresponding detectors.

Method	Detectors	mAP	AP50	AP75	APS	APM	APL	#Params	FLOPs
SAM-Huge	Faster-RCNN (R50)	35.6	54.9	38.4	17.2	39.1	51.4	641M	3,001G
SAM-Huge (finetuned)	Faster-RCNN (R50)	35.8	55.0	38.4	16.5	38.6	53.0	641M	3,001G
Open-Vocabulary SAM	Faster-RCNN (R50)	35.8	55.6	38.3	16.0	38.9	53.1	304M	1,180G
SAM-Huge	Detic (swin-base)	36.4	57.1	39.4	21.4	40.8	54.6	641M	3,001G
SAM-Huge (finetuned)	Detic (swin-base)	36.8	57.4	39.8	20.8	40.6	55.1	641M	3,001G
Open-Vocabulary SAM	Detic (swin-base)	36.7	57.2	39.7	20.7	40.8	54.9	304M	1,180G
SAM-Huge	ViTDet (Huge)	46.3	72.0	49.8	25.2	45.5	59.6	641M	3,001G
SAM-Huge (finetuned)	ViTDet (Huge)	46.5	72.3	50.3	25.2	45.8	60.1	641M	3,001G
Open-Vocabulary SAM	ViTDet (Huge)	48.8	73.8	52.9	24.8	46.3	64.2	304M	1,180G

Method	1-IoU (COCO)	cls.	open.
SAM [30] (H)	78.2	-	-
SEEM [91] (T)	73.7	✓	-
Semantic-SAM [32] (T)	76.1	✓	-
OV-SAM (ours)	81.7	✓	✓

- 与SAM表现对比

不论是使用detector还是交互式提示，都比SAM好

Table 5: Scaling up with large-scale datasets.

Datasets	Accuracy	#vocaulary	#images
LVIS	83.1	1,203	99K
V3Det	78.7	13,204	183K
I-21k	44.5	19,167	13M
V3Det + LVIS	82.7	13,844	282K
V3Det + LVIS + I-21k	83.3	25,898	13M
V3Det + LVIS + I-21k + Object365	83.0	25,970	15M

Backbone	IoU	Acc	#FLOPs(G)	#Params (M)
RN50	77.3	50.8	728	165
RN50x16	78.1	55.1	1,180	304
RN50x64	78.1	54.1	2,098	568
ConvNeXt-L	78.3	59.1	1,313	321
ViT-L-14	38.6	14.3	2,294	441

Table S1: Comparison with Recent Joint SAM and CLIP models [6, 20, 59, 83].

Property	SSA [6]	SAM-CLIP [59]	RecognizeAnything [83]	Sambor [20]	Ours
Single Backbone	✗	✓	✗	✗	✓
Object-Level Classification	✗	✗	✗	✓	✓
Interactive Segmentation	✗	✓	✗	✗	✓

- 扩大数据集

通过交互式提示，能识别超过22000个类

- 更换CLIP的backbone

基于卷积的backbone表现更好

- 与最近的结合SAM和CLIP的模型比较

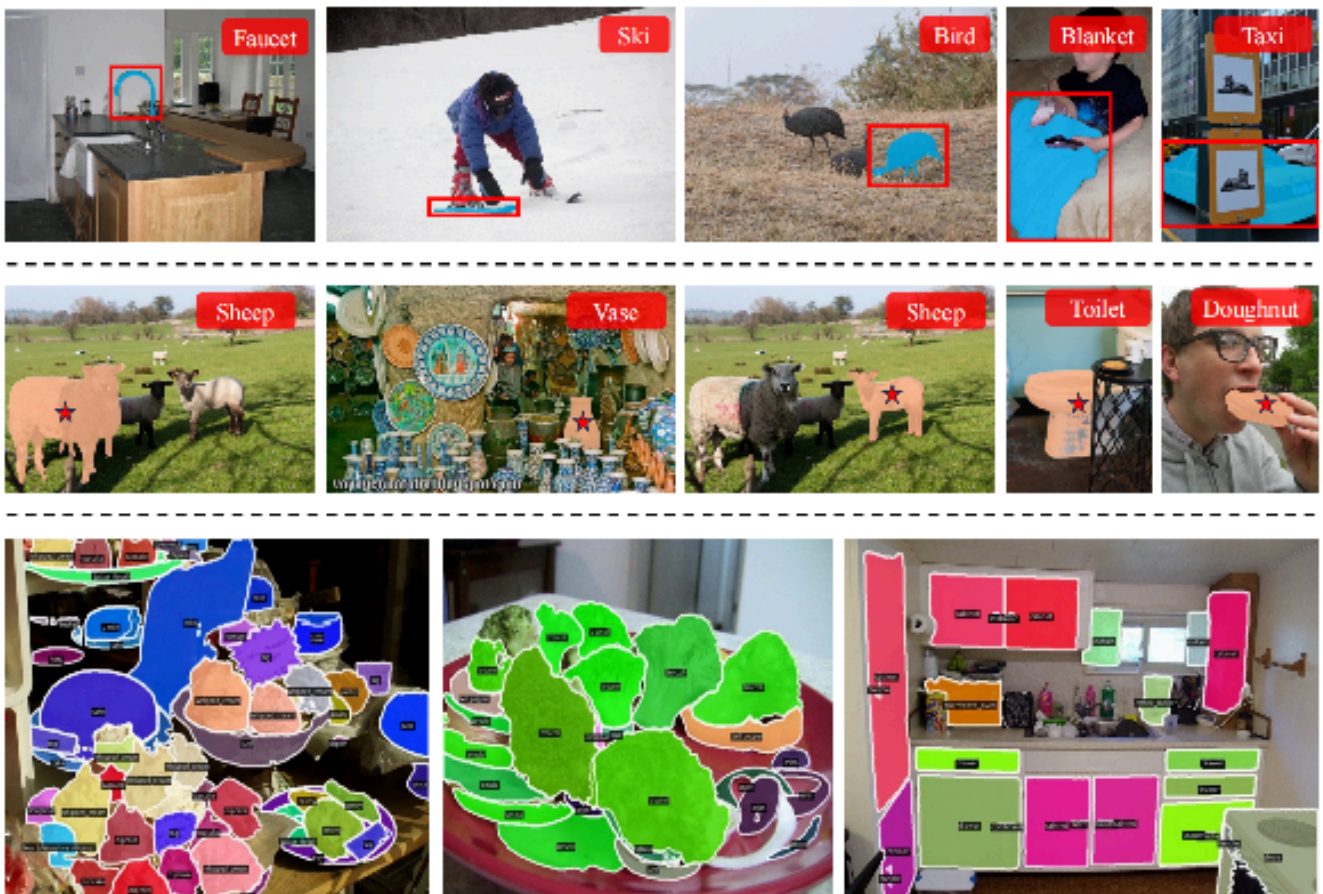

Faucet
Ski
Bird
Blanket
Taxi
Sheep
Vase
Sheep
Toilet
Doughnut

Fig. S3: Visualization of everything mode.

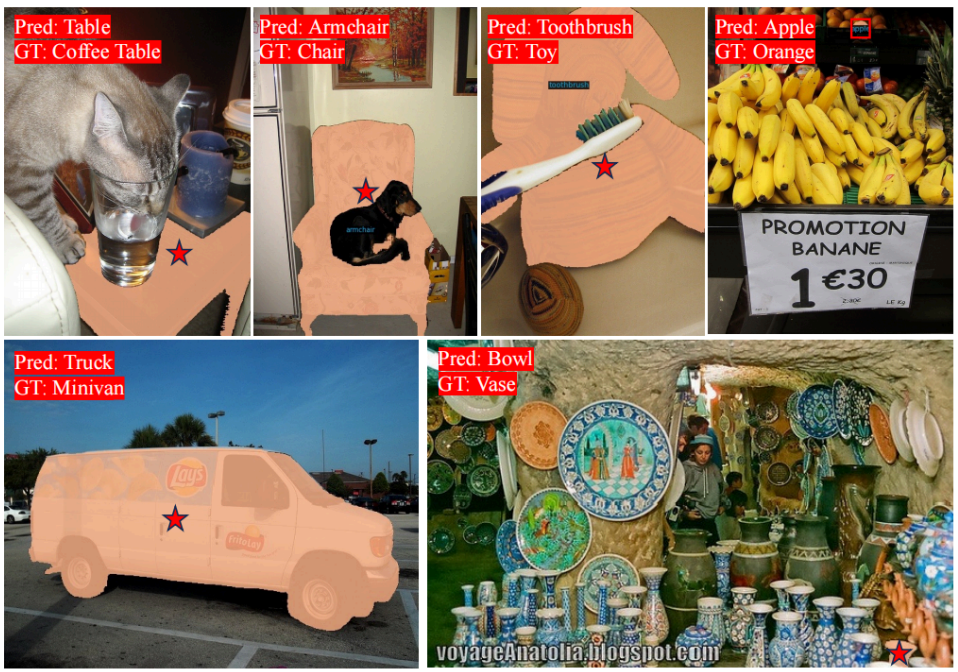


Fig. S4: Failure cases of Open-Vocabulary SAM.