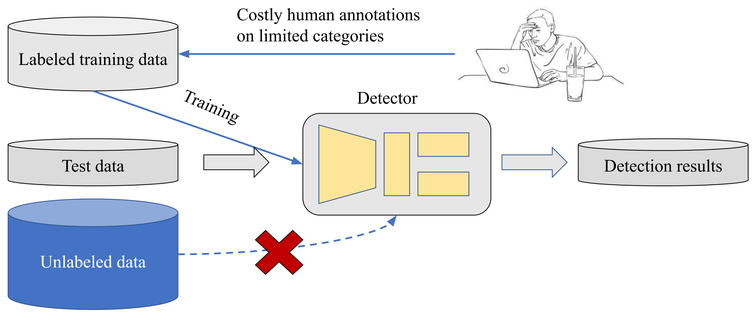
Exploiting Unlabeled Data with Vision and Language Models for Object Detection

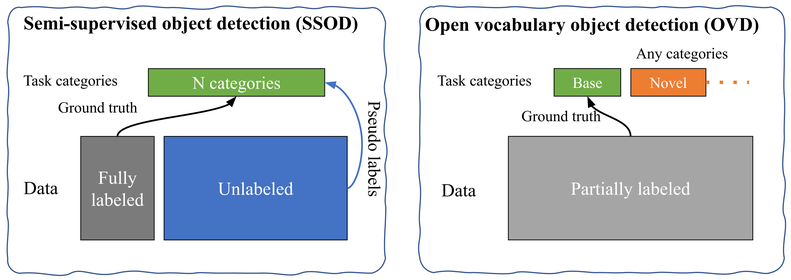
Shiyu Zhao[1], Zhixing Zhang[1], Samuel Schulter[2], Long Zhao[3],
Vijay Kumar B.G[2], Anastasis Stathopoulos[1], Manmohan Chandraker[2,4], Dimitris Metaxas[1]
[1]Rutgers University [2]NEC Laboratories America [3]Google Research [4]UC San Diego

Introduction

❖ Drawbacks of traditional object detection training: **Limited by costly human annotations & unable to leverage unlabeled data**.



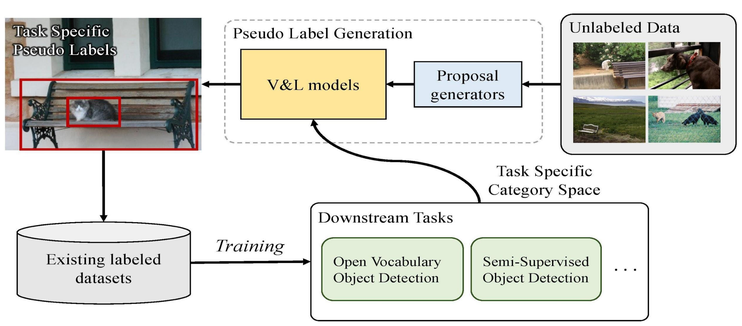
❖ Two approaches to leverage unlabeled data.



😕 Fixed & limited categories 😄 Any/infinite categories

😄 Unlabeled images used 😕 Unlabeled images not leveraged

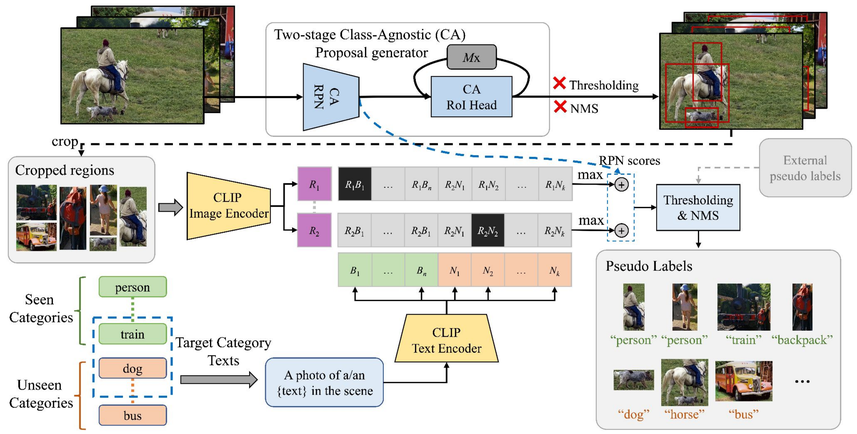
❖ This paper propose VL-PLM which can use unlabeled images and have a flexible label space.



😄 Unlabeled images used 😄 Any/infinite categories

Approach

❖ VL-PLM contains three steps, 1) generate region proposals using a pretrained two-stage class-agnostic proposal generator, 2) classify region proposals into categories of interests with pretrained V&L model (CLIP), and 3) train a detector with the pseudo labels.



Key Results

❖ Quantitative results

Method	Training data	AP_r	AP_c	AP_f	mAP
Supervised	Base + Novel	12.3	**24.3**	32.4	25.4
ViLD [16]	Base	16.6	21.1	31.6	24.4
VL-PLM (Ours)	Base	**17.2**	23.7	**35.1**	**27.0**

OVD on LVIS

Methods	1% COCO	2% COCO	5% COCO	10% COCO
Supervised	9.25	12.70	17.71	22.10
Supervised+PLs	11.18	14.88	21.20	25.98
Supervised+VL-PLM	**15.35**	**18.60**	**23.70**	**27.23**
STAC [46]	13.97	18.25	24.38	28.64
STAC+VL-PLM	**17.71**	**21.20**	**26.21**	**29.61**

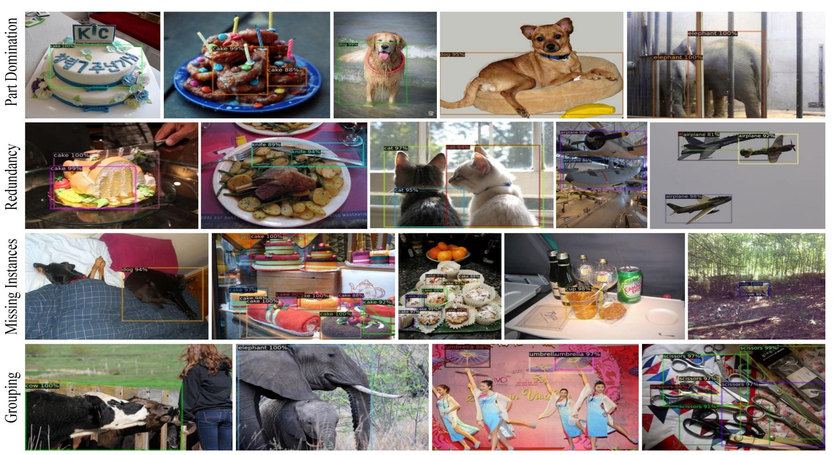
SSOD on COCO

Method	Training Source	Novel AP	Base AP	Overall AP
Bansal *et al.* [4]		0.31	29.2	24.9
Zhu *et al.* [63]	instance-level labels in $\mathcal{S}_B$	3.41	13.8	13.0
Rahman *et al.* [40]		4.12	35.9	27.9
OVR-CNN [56]	image-caption pairs in $\mathcal{S}_B \cup \mathcal{S}_N$ instance-level labels in $\mathcal{S}_B$	22.8	46.0	39.9
Gao *et al.* [14]	raw image-text pairs via Internet	30.8	46.1	42.1
RegionCLIP [59]	image-caption pairs in $\mathcal{S}_B \cup \mathcal{S}_N$ instance-level labels in $\mathcal{S}_B$	31.4	57.1	50.4
RegionCLIP* [59]	raw image-text pairs via Internet	14.2	52.8	42.7
ViLD [16]	instance-level labels in $\mathcal{S}_B$	27.6	59.5	51.3
VL-PLM (Ours)		**34.4**	**60.2**	**53.5**

Zero-shot/OVD on COCO

❖ Visualizations

Visualization of good pseudo labels



Visualization of bad pseudo labels