OADP

Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection

Luting Wang[1,3] Yi Liu[1,3] Penghui Du[1,3] Zihan Ding[1,3] Yue Liao[1,3]* Qiaosong Qi[2]
Biaolong Chen[2] Si Liu[1,3]

[1]IAI, BUAA [2]Alibaba [3]Hii, BUAA

WED-AM-281

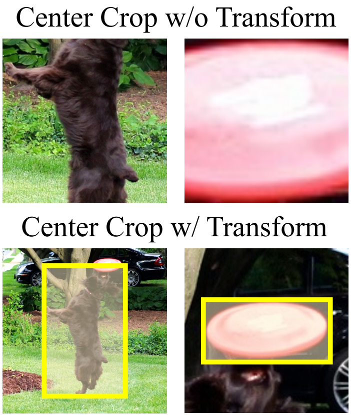
Table of Contents

1. Abstract
2. Introduction
3. Method
4. Comparisons
5. Ablation Study

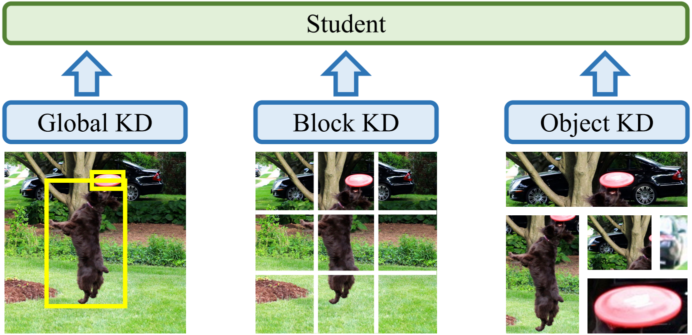
Abstract

Motivation

Knowledge Extraction



Knowledge Transfer



Abstract

Object-Aware Distillation Pyramid

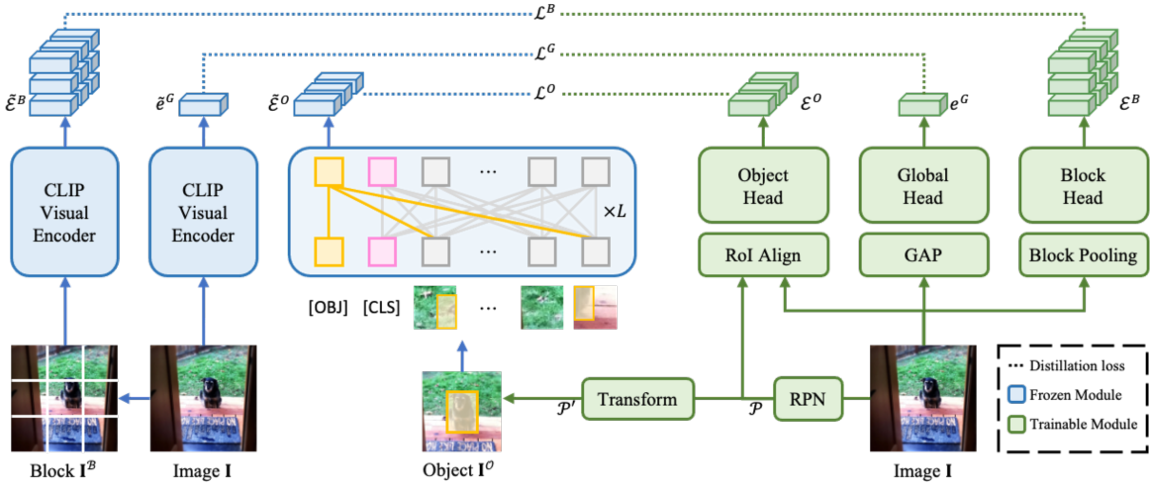
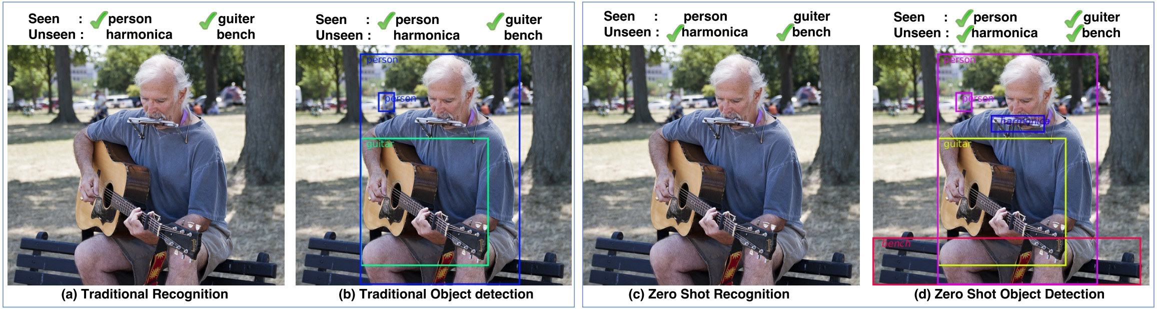


Table of Contents

1. Abstract
2. Introduction
3. Method
4. Comparisons
5. Ablation Study

From Closed-Set to Open-Set

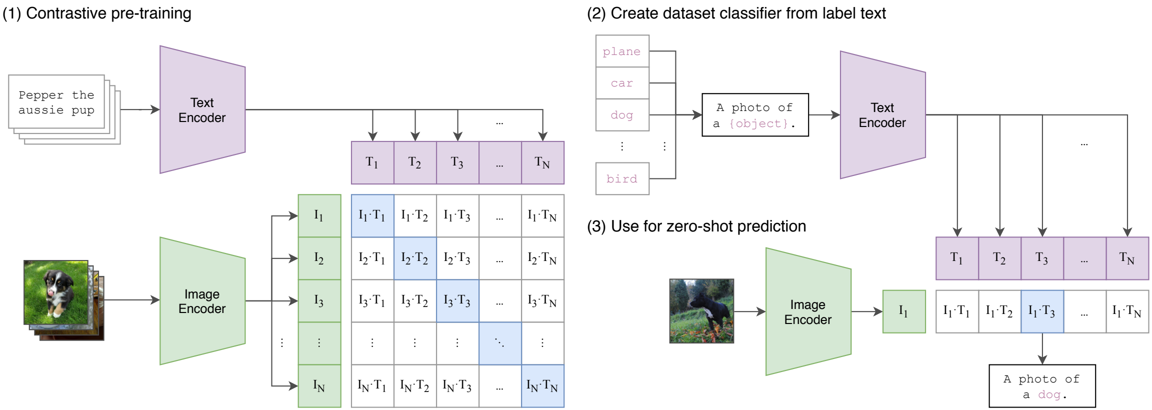
- Most object detectors recognize only known objects.
- Real-world applications require detectors that can detect unknown objects.
- Zero-shot detectors can recognize and locate novel objects without annotations.



(a) Traditional Recognition (b) Traditional Object detection (c) Zero Shot Recognition (d) Zero Shot Object Detection

Rahman, Shafin, *et al.* "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts." ACCV. 2019.

CLIP

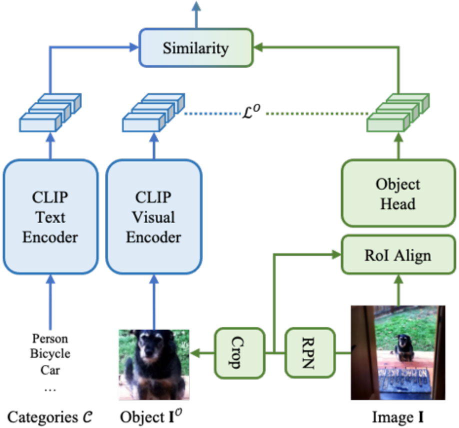

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Radford, Alec, *et al.* “Learning transferable visual models from natural language supervision.” ICML. 2021.

Open-Vocabulary Object Detection

- CLIP text encoder extracts generalizable category embeddings for open-vocabulary classification.
- CLIP visual encoder guides the object detector to learn better visual features.
- CLIP-guided detectors belong to open-vocabulary object detection (OVD).



Benchmarks

According to the training data, we summarize the existing OVD methods into four types:

	V-OVD	C-OVD	G-OVD	WS-OVD
Image Caption A leaping dog.		✓		✓
Category Prior Novels: dog, . . .			✓	✓
Image Label frisbee, dog, . . .				✓
Representative	ViLD	OVR-CNN	VL-PLM	Detic

Gu, Xiuye, *et al*. "Open-vocabulary object detection via vision and language knowledge distillation." ICLR. 2021.

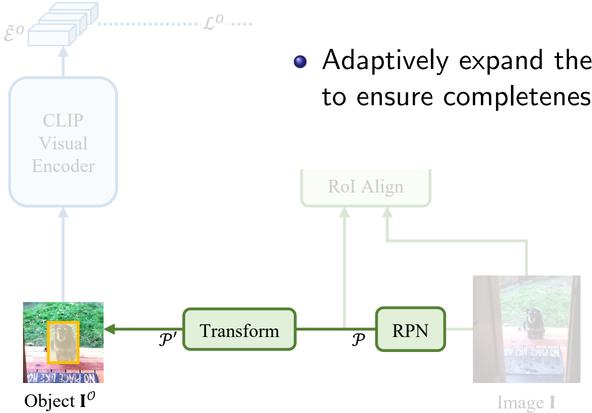
Zareian, Alireza, *et al*. "Open-vocabulary object detection using captions." CVPR. 2021.

Zhao, Shiyu, *et al*. "Exploiting unlabeled data with vision and language models for object detection." ECCV. 2022.

Zhou, Xingyi, *et al*. "Detecting twenty-thousand classes using image-level supervision." ECCV. 2022.

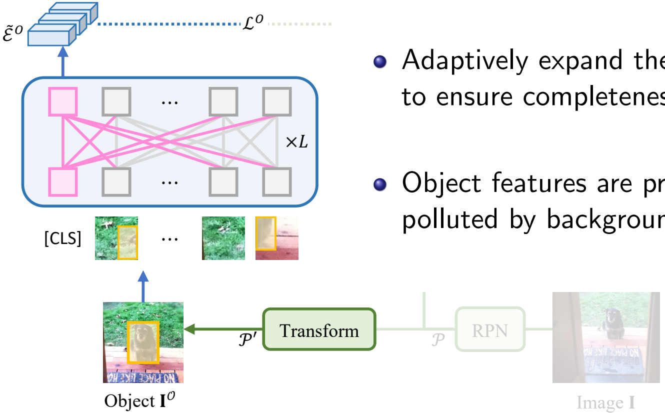
Table of Contents

Object-Aware Knowledge Extraction



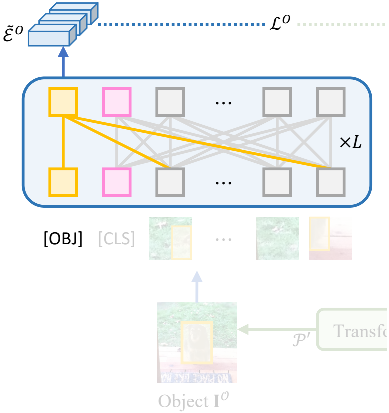
- Adaptively expand the proposals to ensure completeness

Object-Aware Knowledge Extraction



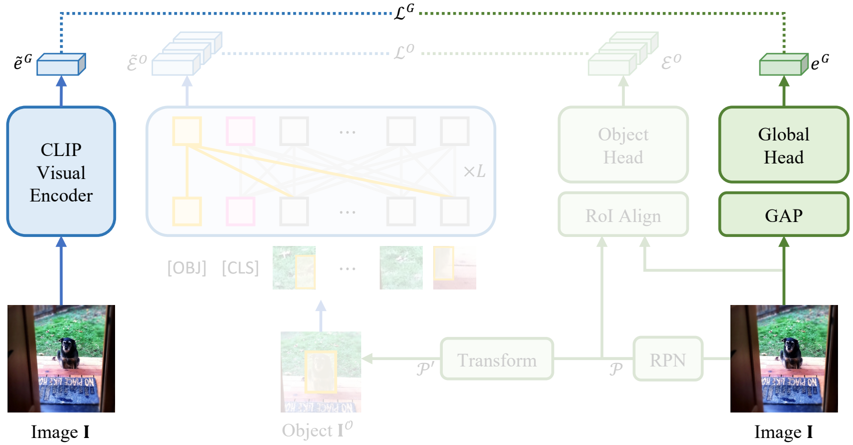
- Adaptively expand the proposals to ensure completeness
- Object features are prone to be polluted by background noise

Object-Aware Knowledge Extraction

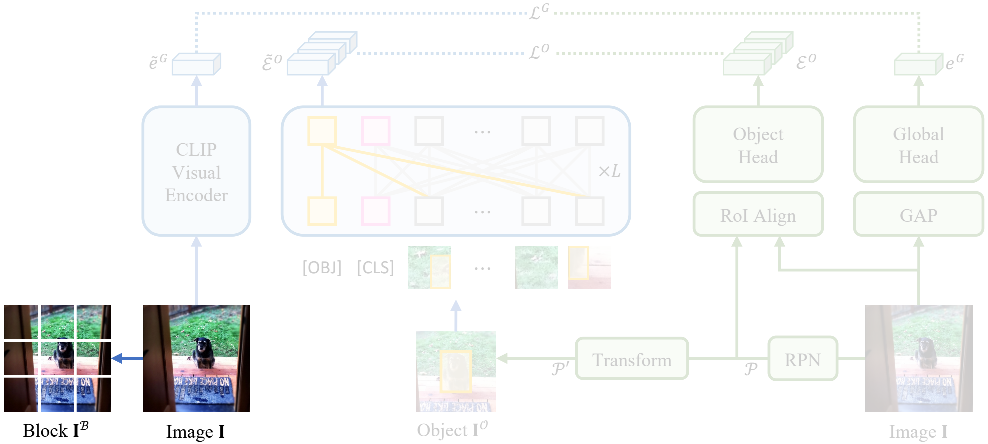


- Adaptively expand the proposals to ensure completeness
- Object features are prone to be polluted by background noise
- Introduce [OBJ] token attending to object regions only

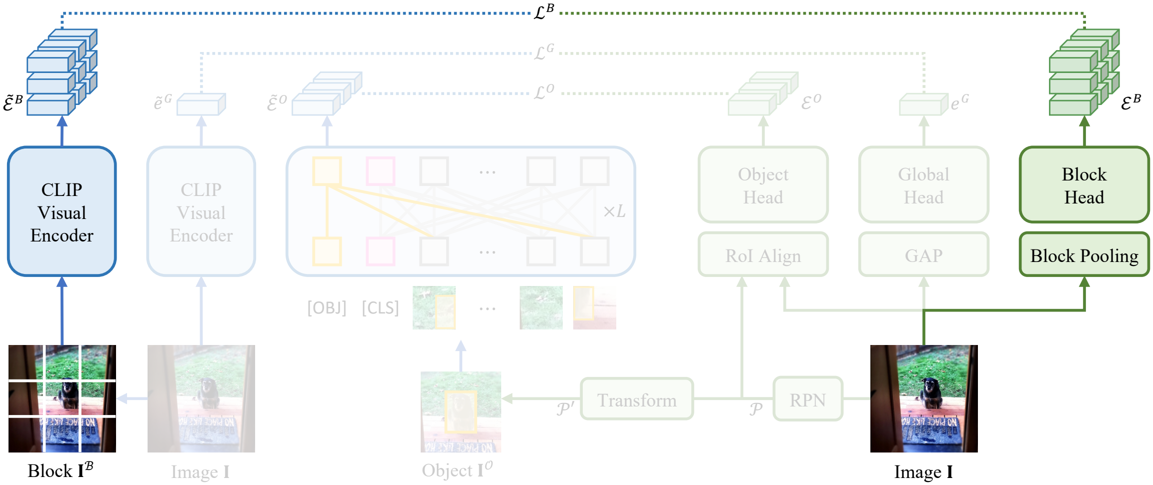
Global Distillation



Block Distillation



Block Distillation



Distillation Pyramid

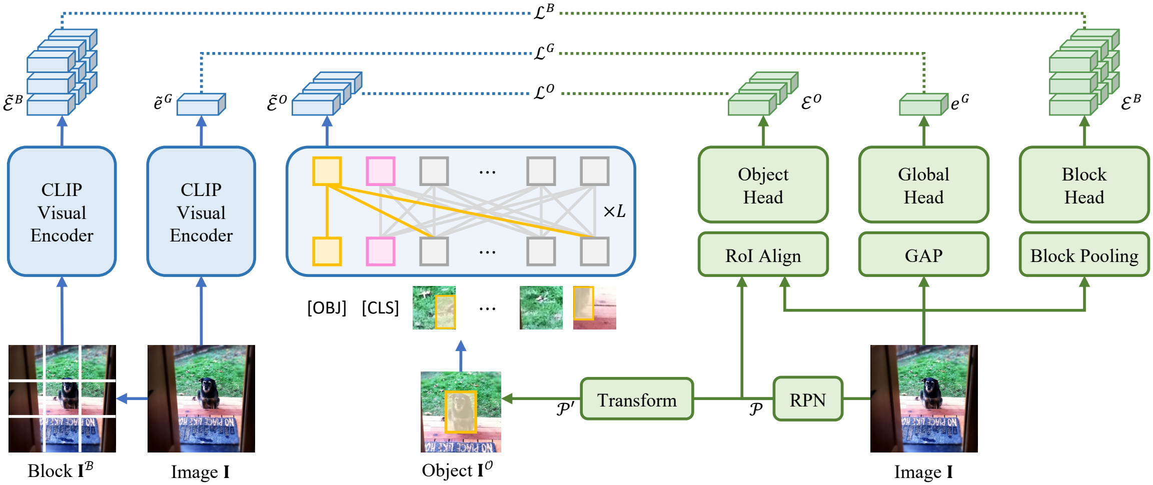


Table of Contents

OV-COCO

- We follow OV-RCNN and divide the MS-COCO 2017 dataset into 48 base categories and 17 novel categories.
- Our OADP achieves state-of-the-art performance on both V-OVD and G-OVD.

Benchmark	Method	mAP$^{N}_{50}$	mAP$^{B}_{50}$	mAP$_{50}$
V-OVD	ViLD	27.6	59.5	51.3
	RegionCLIP*	14.2	52.8	42.7
	OADP (Ours)	**30.0**	53.3	47.2
C-OVD	OVR-CNN	22.8	46.0	39.9
	HierKD	20.3	51.3	43.2
	RegionCLIP	26.8	54.8	47.5
	LocOV	28.6	51.3	45.7
	PB-OVD	29.1	44.4	40.4
G-OVD	OV-DETR	29.4	61.0	52.7
	VL-PLM	32.3	54.0	48.3
	OADP (Ours)	**35.6**	55.8	50.5
WS-OVD	Detic	27.8	47.1	45.0

Zareian, Alireza, *et al.* "Open-vocabulary object detection using captions." CVPR. 2021.

OV-LVIS

- Some experiments are conducted under the OV-LVIS setting, where the 337 rare categories in LVIS are treated as novel categories, and the other 866 are base categories.
- Metrics for the OV-LVIS setting are $\mathrm{AP_r}$, $\mathrm{AP_c}$, $\mathrm{AP_f}$, and AP.
- Both object detection and instance segmentation metrics are reported.

Method	Object Detection				Instance Segmentation			
	$\mathrm{AP_r}$	$\mathrm{AP_c}$	$\mathrm{AP_f}$	AP	$\mathrm{AP_r}$	$\mathrm{AP_c}$	$\mathrm{AP_f}$	AP
ViLD	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
DetPro	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OV-DETR	-	-	-	-	17.4	25.0	32.5	26.6
OADP (Ours)	**21.9**	28.4	32.0	28.7	**21.7**	26.3	29.0	26.6

Table of Contents

1 Abstract

2 Introduction

3 Method

4 Comparisons

5 Ablation Study

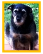
Object-Aware Distillation Pyramid

- We conduct ablation studies on the distillation modules in OADP.
- The baseline is our re-implemented ViLD-ensemble model.

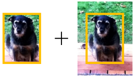
Global	Block	Object	$\text{mAP}^{\text{N}}_{50}$	$\text{mAP}^{\text{B}}_{50}$	mAP_{50}
			24.99	50.29	43.67
✓			25.72	51.89	45.04
	✓		27.25	53.60	46.71
		✓	27.23	55.96	48.45
✓	✓		26.49	51.25	44.78
✓		✓	28.80	54.29	47.62
	✓	✓	29.01	55.45	48.53
✓	✓	✓	**29.95**	53.26	47.17

Object-Aware Knowledge Extraction

Original

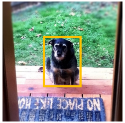


Baseline

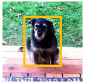


ViLD*

MBS



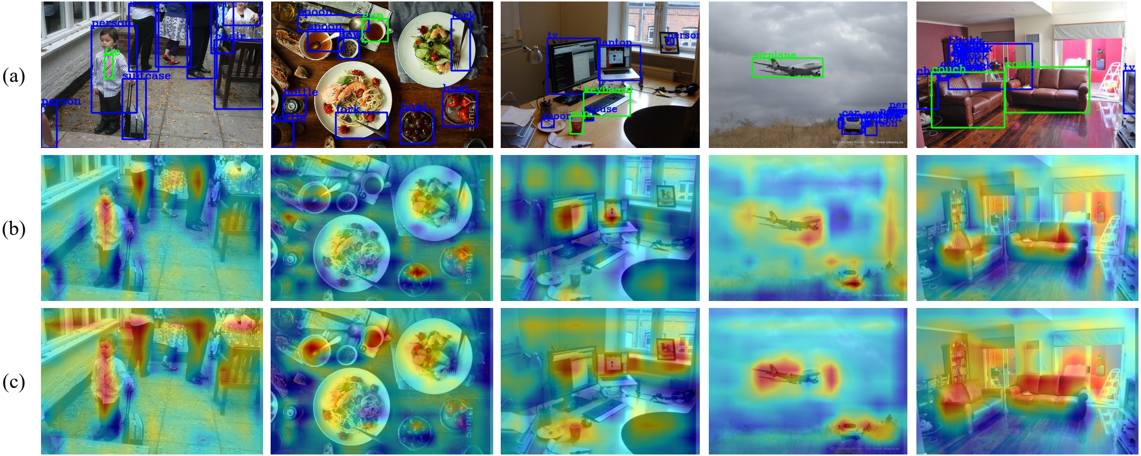
Fixed



Adaptive

Method	Macro Precision		Weighted Precision	
	w/o OAKE	w/ OAKE	w/o OAKE	w/ OAKE
Baseline	58.08	-	62.04	-
ViLD*	63.36	-	65.91	-
MBS	61.70	63.83	64.81	65.82
Fixed	49.07	64.53	51.49	**69.75**
Adaptive	51.64	**66.09**	55.85	68.68

Visualization



OADP

Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection

CONTACT US

arxiv.org/abs/2303.05892

github.com/LutingWang/OADP

@ liaoyue.ai@gmail.com