What Makes Good Open-Vocabulary Detector: A Disassembling Perspective

Qihoo 360 AI Research

Jincheng Li, Chunyu Xie, Xiaoyu Wu, Bin Wang, Dawei Leng

Outline

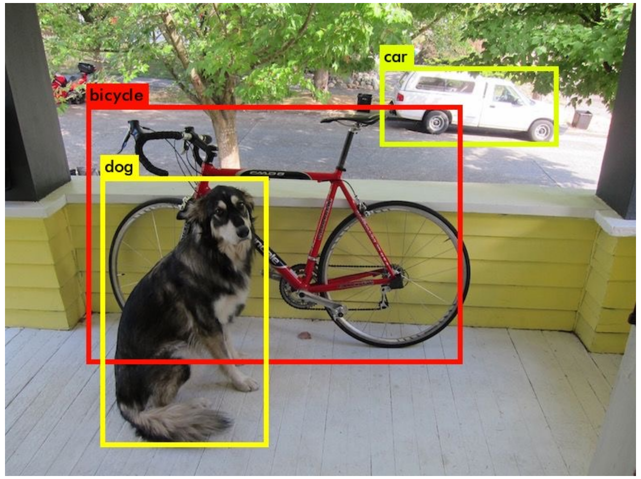
- Traditional object detection
- Open-vocabulary object detection
- Our approach
- Experiment
- Conclusion

Traditional object detection

bicycle | dog | car

Train

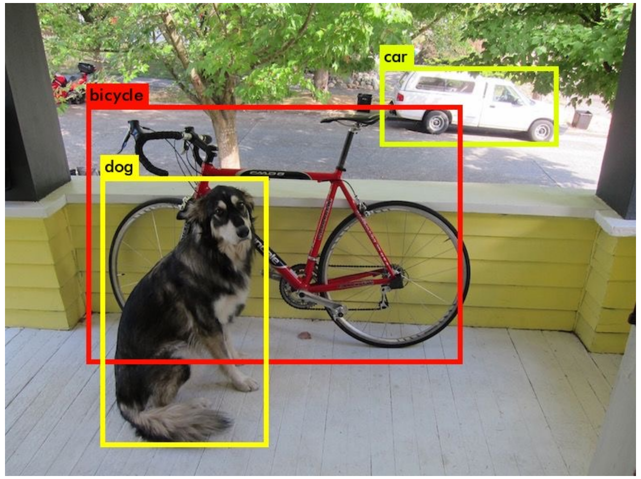
Test



Traditional object detection:

- require a variety of fine-grained annotations: ***bicycle, dog, car***
- localize and classify seen categories: ***bicycle, dog, car***
- human annotations are costly and tedious

Open-vocabulary object detection (OVD)

	bicycle	dog	car
Train		×	×
Test			

Open-vocabulary object detection:

- require small pre-defined (base) categories: ***bicycle***
- predict pre-defined and unseen (novel) categories: ***bicycle, dog, car***

Challenges of OVD

- Whether to use detection-tailored pre-trained CLIP remains an open question
- How to effectively improve the detection ability under the settings of the OVD task is still a challenge

Our goal

- We set out to address these issues under the OVD settings
- Our goal is to analyze which part of localization and classification can improve the overall performance of OVD task

Three families of OVD methods

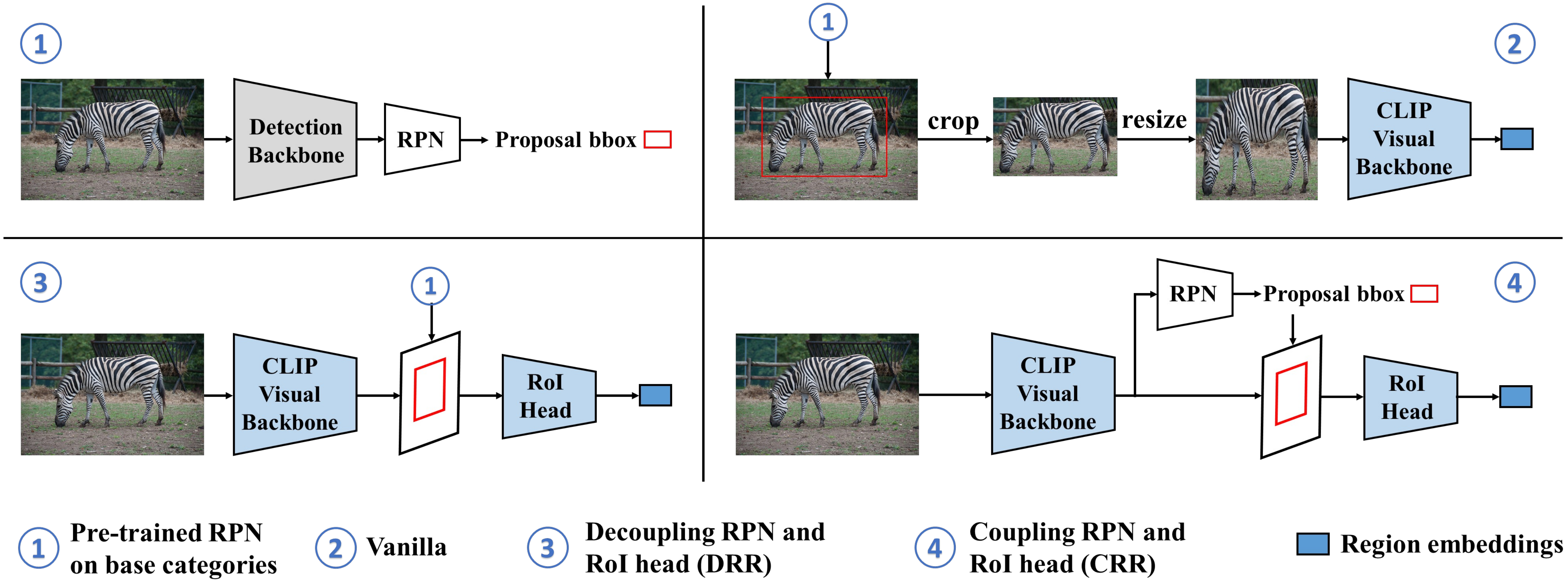


Figure 1: An overview of three approaches: a vanilla method, DRR and CRR.

Experiment

Table 1: Performance on OVD-COCO compared with state-of-the-art methods.

Method	Extra Dataset	Backbone	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
OVR-CNN [24]	COCO Captions	ResNet50	22.8	46.0	39.9
ViLD [7]	-	ResNet50	27.6	59.5	51.2
Detic [28]	COCO Captions	ResNet50	27.8	47.1	45.0
RegionCLIP [26]	CC3M	ResNet50	31.4	57.1	50.4
BARON [20]	COCO Captions	ResNet50	33.1	54.8	49.1
Vanilla (Ours)	-	ResNet50	31.8	37.2	35.5
CRR (Ours)	CC3M	ResNet50	32.0	52.5	47.1
DRR (Ours)	CC3M	ResNet50	**35.8**	54.6	49.6

- **The vanilla method achieves comparable results on novel categories but obtains bad results on base categories**
- **CRR obtains a higher Novel AP than RegionCLIP, but lower than BARON**
- **DRR achieves the best results and outperforms BARON by 2.7 Novel AP**

Experiment

Table 2: Performance on OVD-LVIS compared with state-of-the-art methods.

Method	Backbone	Require Novel Class [20]	AP_r	AP_c	AP_f	mAP
RegionCLIP [26]	ResNet50	×	17.1	27.4	34.0	28.2
ViLD [7]	ResNet50	✓	16.7	26.5	34.2	27.8
BARON [20]	ResNet50	✓	20.1	28.4	32.2	28.4
Vanilla (Ours)	ResNet50	×	17.2	14.8	11.5	13.9
CRR (Ours)	ResNet50	×	14.0	23.7	28.5	21.9
DRR (Ours)	ResNet50	×	20.1	29.9	35.7	30.5
DRR (Ours)	ResNet50	✓	**22.0**	25.4	33.7	28.1

- **DRR achieves 20.1 AP_r, which is significantly better than RegionCLIP by 3 AP_r**
- **Similar to OVD-COCO, CRR still leads a competitive result**
- **The vanilla method obtains bad results compared to other methods**

Experiments of vanilla method

Table 3: Influence of object localization on OVD-COCO. Faster

Table 5: Effect of image embedding ensemble on OVD-COCO.

Method	Ensemble	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
Vanilla	×	27.3	28.9	28.5
Vanilla	✓	29.6	32.1	31.1

Experiments of DRR

Table 7: Effect of CLIP visual backbone on OVD-COCO compared with state-of-the-art methods.

Method	Visual Backbone	Detection-tailored Pre-training	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
RegionCLIP [26]	ResNet50	×	14.2	52.8	42.7
RegionCLIP [26]	ResNet50	✓	31.4	57.1	50.4
DRR (Ours)	ResNet50	✓	**35.8**	54.6	49.4
RegionCLIP [26]	ResNet50x4	✓	39.3	61.6	55.7
DRR (Ours)	ResNet50x4	✓	**41.9**	57.8	53.7

Table 6:

- **Replacing RPN with Faster R-CNN cannot achieve the expected results**
- **The significant objectness logits within a better offline RPN are indeed important for model performance**

Table 7:

- **DRR surpasses the previous state-of-the-art (RegionCLIP) by 2.6 AP50 in novel categories**

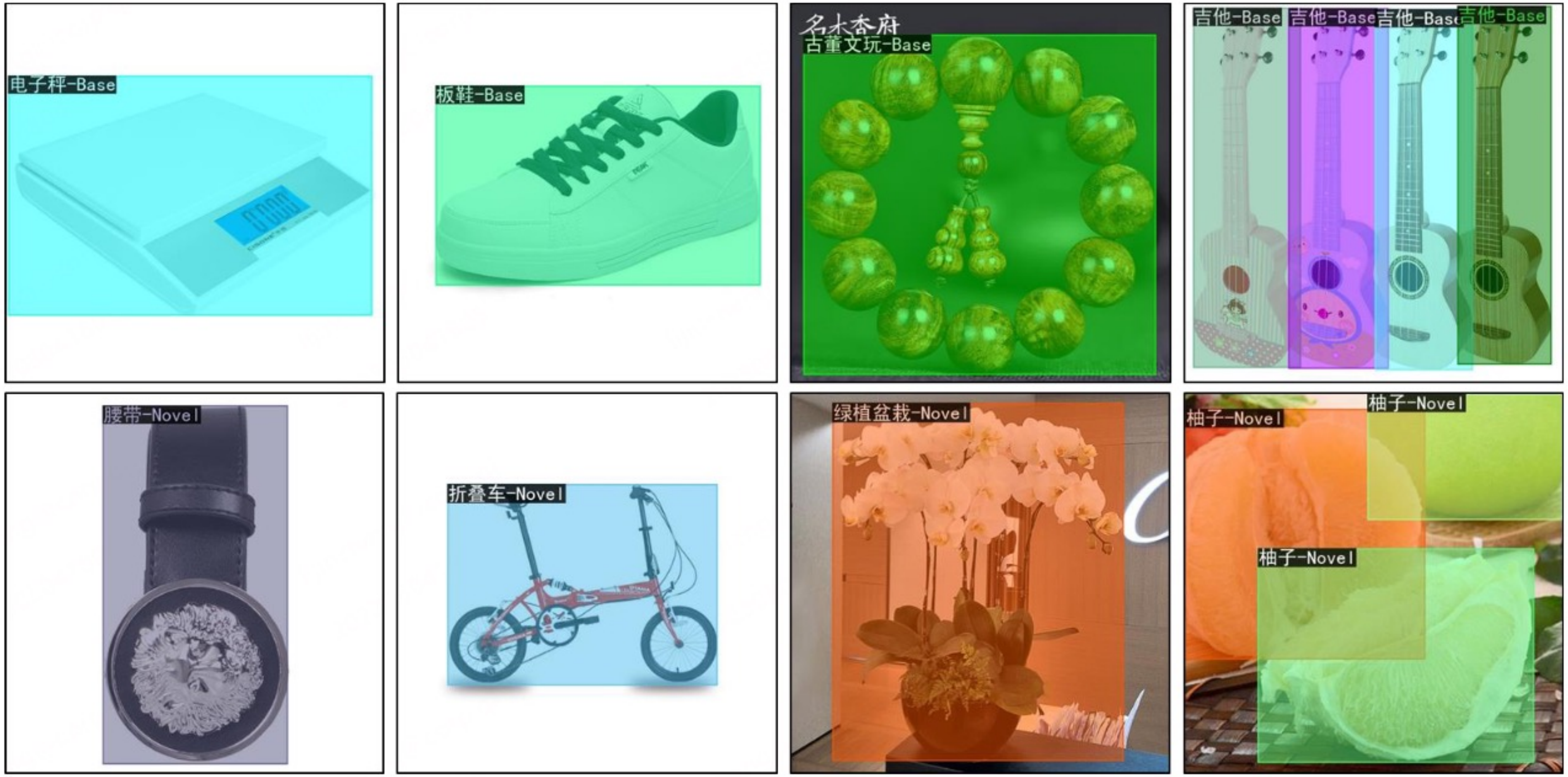
Computational efficiency

Table 8: Comparisons of the computational efficiency over ResNet50 on OVD-COCO.

Method	Params ↓	FPS (BS1 A100) ↑
Vanilla	136.9 M	2
DRR	143.4 M	12
CRR	**111.6 M**	**13**

- **Sharing the visual backbone (CRR) is indeed more effective in specific real-world scenarios**

Product Image Dataset



电子秤: electronic scale
腰带: belt

板鞋: board shoes
折叠自行车: folding bicycle

古董文玩: antique
绿植盆栽: green plants

吉他: guitar
柚子: grapefruit

Figure 2: Examples (with annotations) of PID. The first and second rows are from the base and novel categories, respectively.

Results on PID

Table 9: Comparisons of different fundamental approaches over ResNet50 on PID. *More analysis can be found in Section 5.2.

Method	Visual Backbone	Generalized (233+233)		
		Novel	Base	Overall
Vanilla*	ResNet50	**42.0**	52.8	47.4
DRR	ResNet50	30.7	35.6	33.2
CRR	ResNet50	27.6	34.3	31.0

Thanks!