Aligning Bag of Regions for Open-Vocabulary Object Detection

Size Wu[1] Wenwei Zhang[1] Sheng Jin[2,3] Wentao Liu[3,4] Chen Change Loy[1*]
[1]S-Lab, Nanyang Technological University [2] The University of Hong Kong
[3] SenseTime Research and Tetras.AI [4] Shanghai AI Laboratory
{size001, wenwei001, ccloy}@ntu.edu.sg {jinsheng, liuwentao}@sensetime.com

Paper Tag: WED-PM-276

Outline

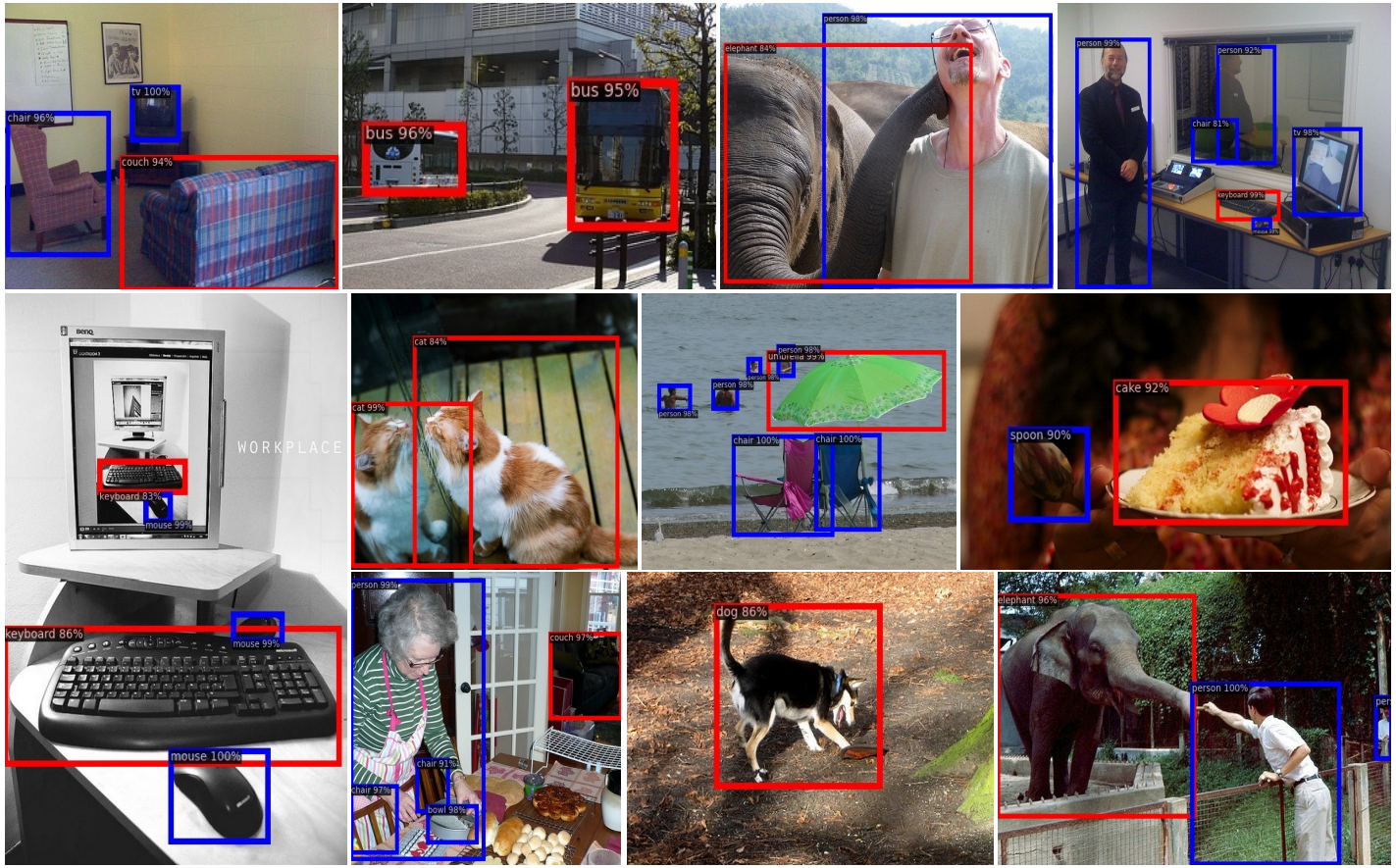
- Introduction
- Method
- Experiment

Outline

- Introduction
- Method
- Experiment

Introduction

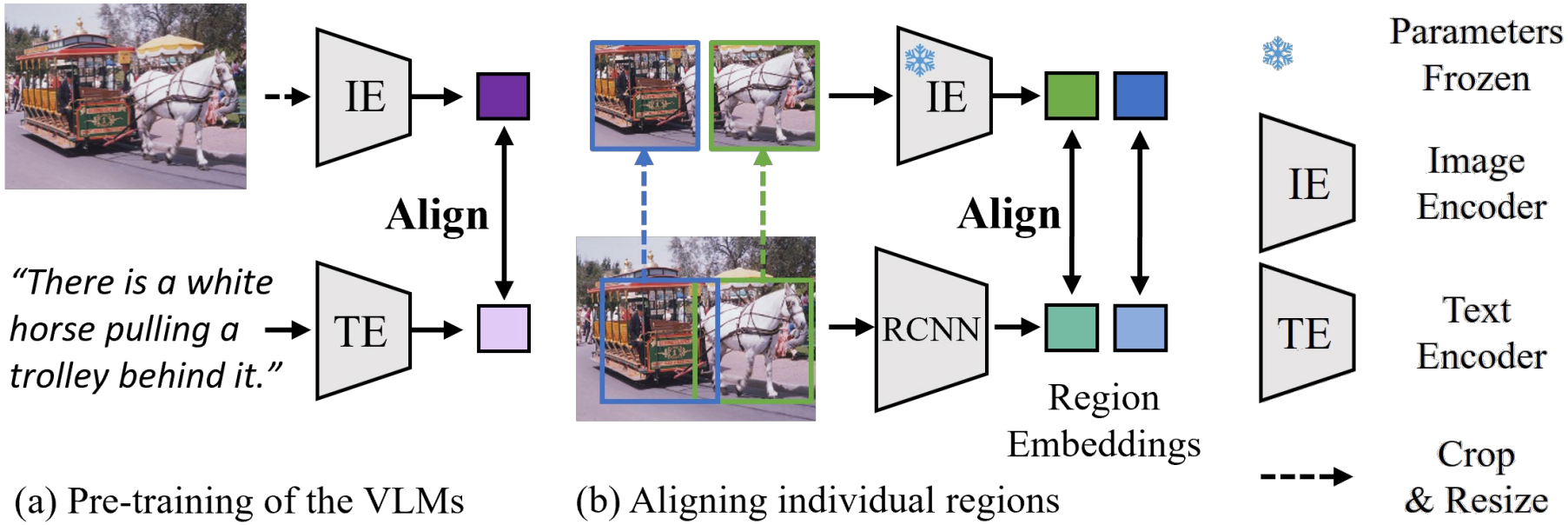
- Open-vocabulary Object Detection



Detecting objects of novel categories unseen in the training phase.

Introduction

- Distillation-based Methods



(a) Pre-training of the VLMs (b) Aligning individual regions

Individually align region embeddings to the corresponding features extracted from the Vison-Language Models (VLMs), e.g., CLIP.

Introduction

- Analysis

"There is a desk." (0.265)

"There is a desk with a monitor." (0.277)

"There is a desk with a monitor and keyboard." (0.283)

"There is a desk with a monitor, keyboard and mouse." (0.294)

"There is a black motorcycle." (0.272)

"There is a black motorcycle parked on the road." (0.279)

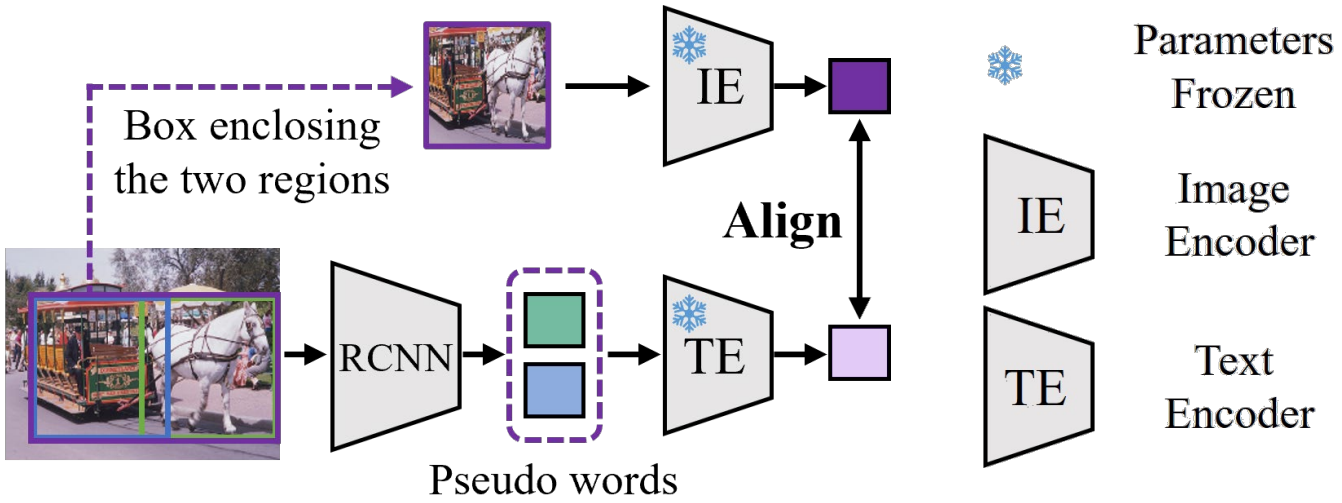
"There are a black motorcycle and a car parked on the road." (0.295)

"There is a black motorcycle parked on the road in front of a car." (0.304)

The VLMs can capture the co-occurrence of objects.

Introduction

- Ours: Aligning Bag of Regions (BARON)



(c) Aligning bag of regions

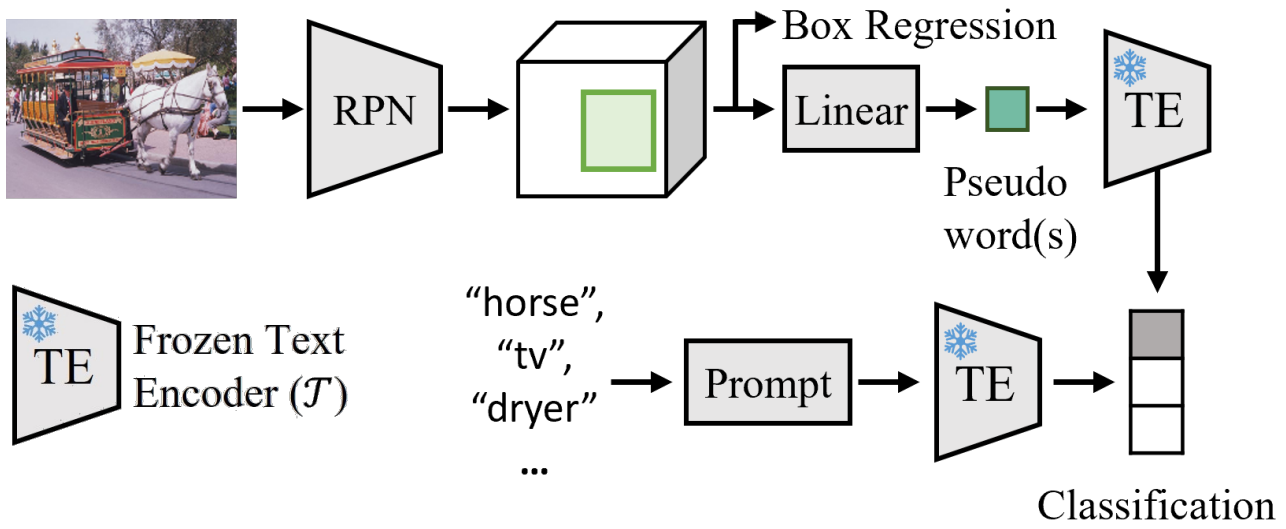
- Regard regions as words
- Mimic the bag-of-words representation of a sentence
- Form a bag of regions to obtain a sentence-like representation

Outline

- Introduction
- Method
- Experiment

Method

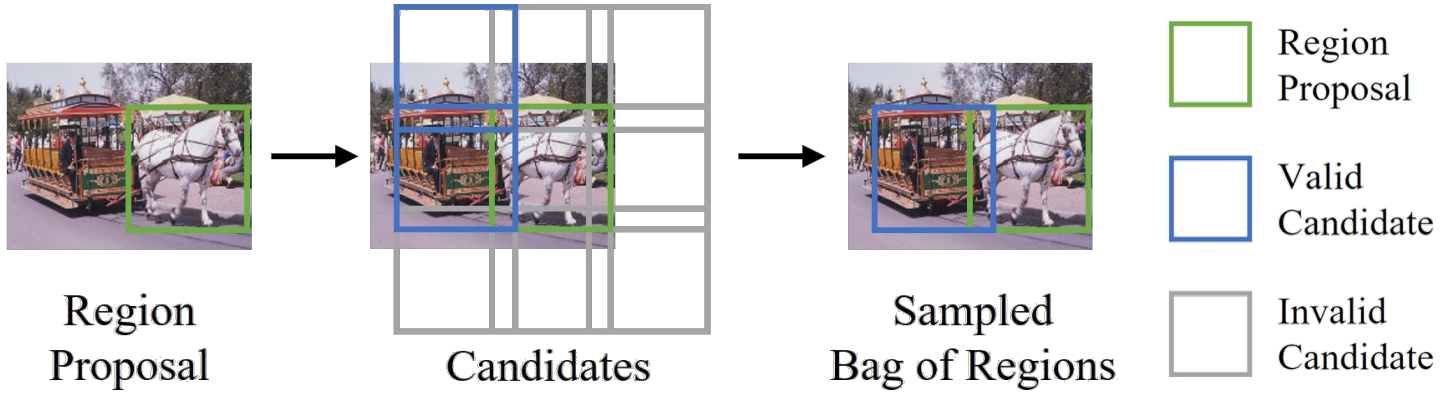
- The Open-vocabulary Detector



For inference and training on base categories

Method

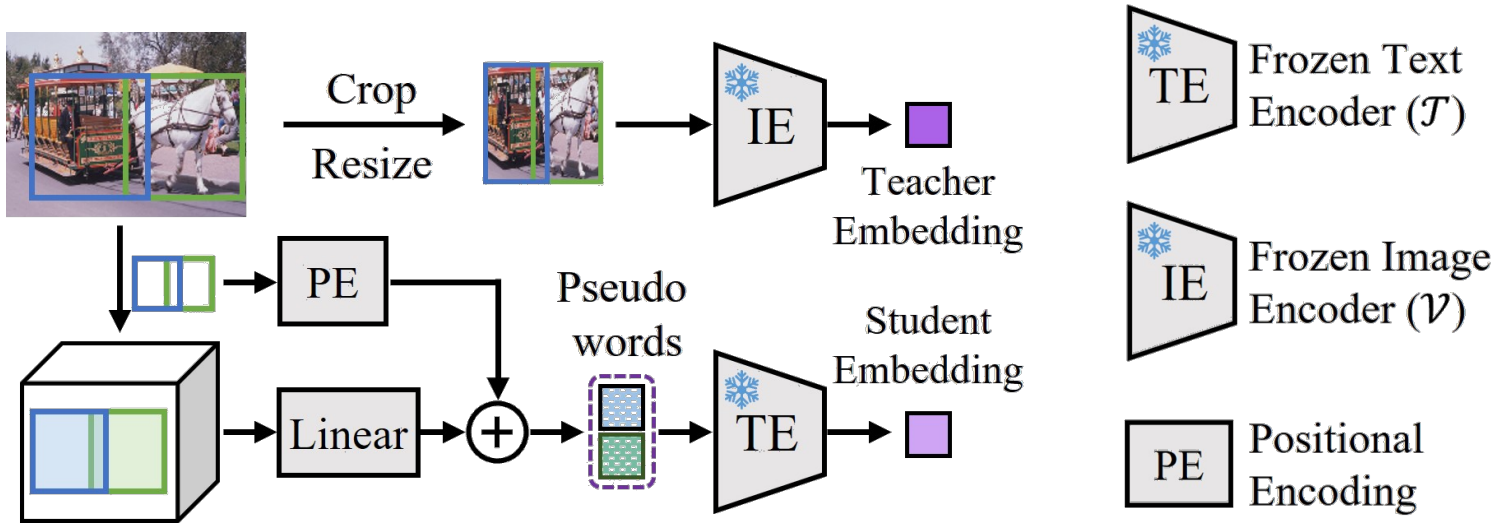
- Forming Bag of Regions



- Start from region proposals
- Sample surrounding (neighboring) region boxes with equal box sizes

Method

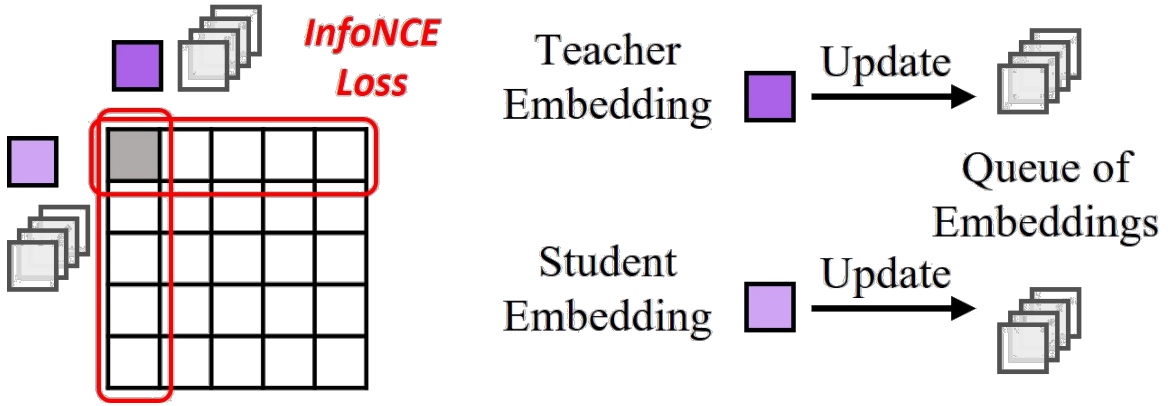
- Representing Bag of Regions



- Student Embedding: Add positional embeddings to the pseudo words, concatenate, and send to the *Text Encoder*
- Teacher Embedding: Send image crop to the *Image Encoder*

Method

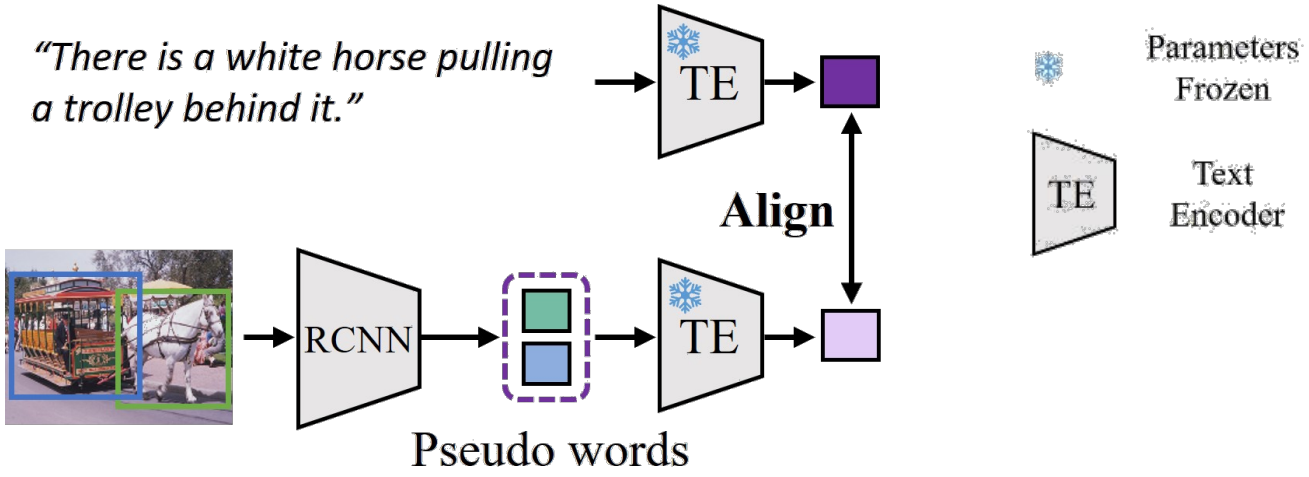
- Aligning Bag of Regions



- Adopt contrastive learning
- Keep queues of embeddings to provide sufficient negative teacher-student embedding pairs

Method

- Caption Supervision



Use the text embedding of image caption as teacher embedding

Outline

- Introduction
- Method
- Experiment

Experiment

- OV-COCO Benchmark

Method	Supervision	Backbone	Detector	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
ViLD [15]	CLIP	ResNet50-FPN	FasterRCNN	27.6	59.5	51.2
OV-DETR [52]	CLIP	ResNet50	DeformableDETR	29.4	61.0	52.7
BARON (Ours)	CLIP	ResNet50-FPN	FasterRCNN	**34.0**	60.4	53.5
OVR-CNN [53]	Caption	ResNet50-C4	FasterRCNN	22.8	46.0	39.9
RegionCLIP [56]	Caption	ResNet50-C4	FasterRCNN	26.8	54.8	47.5
Detic [58]	Caption	ResNet50-C4	FasterRCNN	27.8	51.1	45.0
PB-OVD [13]	Caption	ResNet50-C4	FasterRCNN	30.8	46.1	42.1
VLDet [28]	Caption	ResNet50-C4	FasterRCNN	32.0	50.6	45.8
BARON (Ours)	Caption	ResNet50-C4	FasterRCNN	**33.1**	54.8	49.1
Rasheed *et al.* [41]†	CLIP + Caption	ResNet50-C4	FasterRCNN	36.6	54.0	49.4
BARON (Ours)†	CLIP + Caption	ResNet50-C4	FasterRCNN	**42.7**	54.9	51.7

Experiment

- OV-LVIS Benchmark

Method	Ensemble	Learned Prompt	Object Detection				Instance segmentation			
			AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [15]	-	-	16.3	21.2	31.6	24.4	16.1	20.0	28.3	22.5
OV-DETR [52]	-	-	-	-	-	-	17.4	25.0	32.5	26.6
BARON (Ours)	-	-	**17.3**	25.6	31.0	26.3	**18.0**	24.4	28.9	25.1
ViLD [15]	✓	-	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
ViLD* [15]	✓	-	17.4	27.5	31.9	27.5	16.8	25.6	28.5	25.2
BARON (Ours)	✓	-	**20.1**	28.4	32.2	28.4	**19.2**	26.8	29.4	26.5
DetPro [10]	✓	✓	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
BARON (Ours)	✓	✓	**23.2**	29.3	32.5	29.5	**22.6**	27.6	29.8	27.6

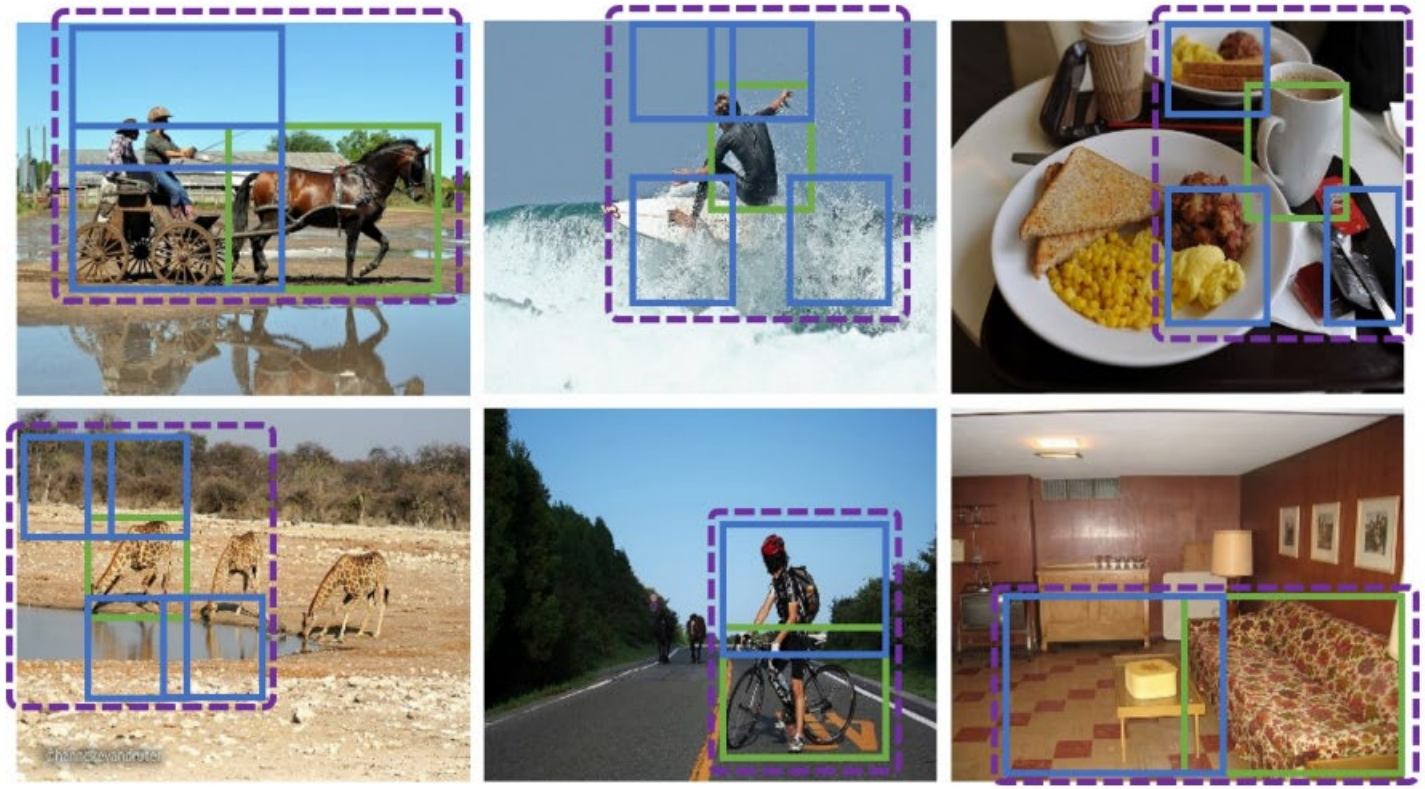
Experiment

- Transfer Results

Method	Pascal VOC		COCO						Objects365					
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised [10]	78.5	49.0	46.5	67.6	50.9	27.1	67.6	77.7	25.6	38.6	28.0	16.0	28.1	36.7
ViLD* [15]	73.9	57.9	34.1	52.3	36.5	21.6	38.9	46.1	11.5	17.8	12.3	4.2	11.1	17.8
BARON (Ours)‡	**74.5**	**57.9**	**36.3**	**56.1**	**39.3**	**25.4**	**39.5**	**48.2**	**13.2**	**20.0**	**14.0**	**4.8**	**12.7**	**20.1**
DetPro [10]	74.6	57.9	34.9	53.8	37.4	22.5	39.6	46.3	12.1	18.8	12.9	4.5	11.5	18.6
BARON (Ours)	**76.0**	**58.2**	**36.2**	**55.7**	**39.1**	**24.8**	**40.2**	**47.3**	**13.6**	**21.0**	**14.5**	**5.0**	**13.1**	**20.7**

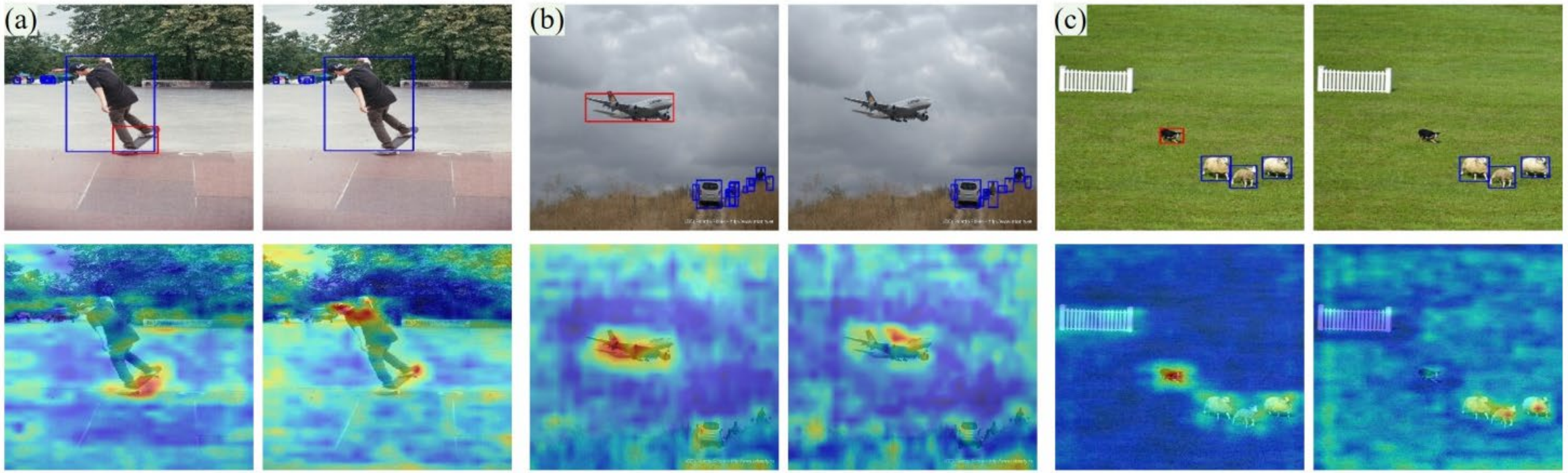
Experiment

- Visualization
  - Bag of Regions



Experiment

- Visualization
  - Featuremap Response

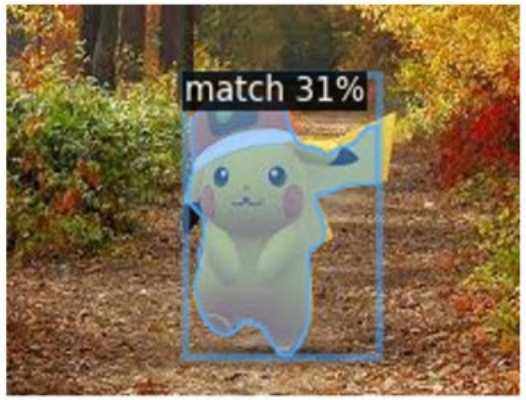


Ours Baseline Ours Baseline Ours Baseline

Experiment

- Visualization
  - Image-based Inference

Reference



Reference

Reference

Reference

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

商汤 sensetime

JUNE 18-22, 2023

CVPR VANCOUVER, CANADA

Thanks for listening!