JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UniDetector

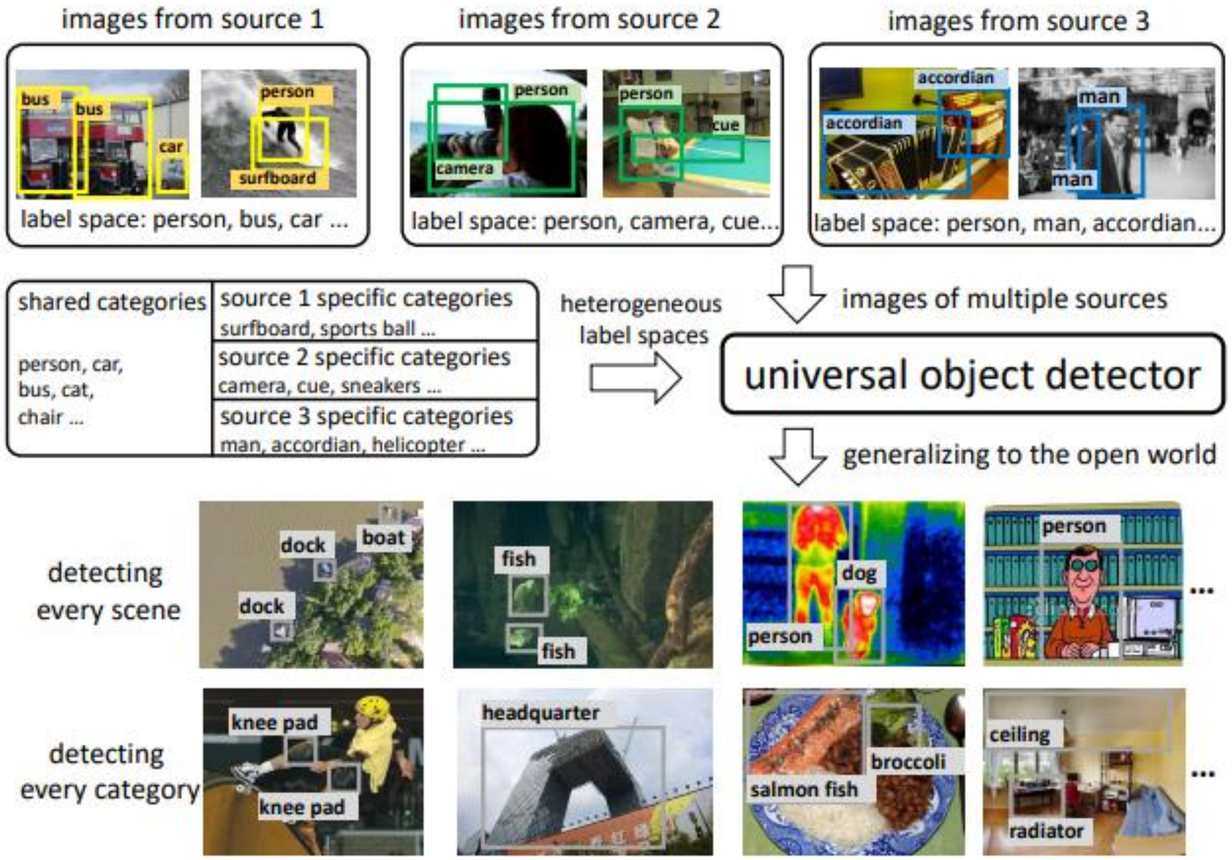
Detecting Everything in the Open World: Towards Universal Object Detection

WED-AM-305 (5082)

Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, Shengjin Wang

Department of Electronic Engineering, Tsinghua University
Beijing National Research Center for Information Science and Technology (BNRist)
The University of Hong Kong Meta Massachusetts Institute of Technology

Overview:



images from source 1
bus
bus
car
person
surfboard
label space: person, bus, car ...
images from source 2
person
camera
person
cue
label space: person, camera, cue...
images from source 3
accordian
accordian
man
man
label space: person, man, accordian...
shared categories
person, car,
bus, cat,
chair ...
source 1 specific categories
surfboard, sports ball ...
source 2 specific categories
camera, cue, sneakers ...
source 3 specific categories
man, accordian, helicopter ...
heterogeneous
label spaces
images of multiple sources
universal object detector
generalizing to the open world
detecting
every scene
dock
boat
dock
fish
fish
dog
person
person
...
detecting
every category
knee pad
knee pad
headquarter
broccoli
salmon fish
ceiling
radiator
...

Previous object detection method pipeline:

Create a detection dataset

- Define the working environment we want
- Define the categories we want to detect (vocabulary)
- Collect images and annotate object instances

ex: common object detection (COCO, Objects365)
traffic detection (Cityscapes, Mapillary Vistas)
UAV (Visdrone)

↓

Training on the dataset

↓

Inference on the dataset

Training on the dataset
Inference on the dataset

The limitation:

focusing on a single dataset

What if we need to detect in a new scene ? (new environment or new categories)

We need to
- collect new images
- annotate again

$\longrightarrow$ Create a new dataset

What we want:

a universal object detector that can detect everything in every scene

once trained, can directly work in unknown situations
without any further re-training

Two abilities that a universal object detector should have:

1. Utilizing images of multiple sources and heterogeneous label spaces for training

a universal object detector that can detect everything in every scene

↓

involving diversified types of images as many as possible

Datasets	Categories	Images
PASCAL VOC	20	11k
Cityscapes	8	5k
MS COCO	80	123k
Objects365	365	638k
LVIS	1230	68k
ImageNet	3130	1.2M
OpenImages	600	1.7M
VisualGenome	80138	108k

Problem:

Limited by human annotators:

1) Large vocabulary datasets are noisy and ambiguous
2) Specialized datasets

Two abilities that a universal object detector should have:

2. Generalizing to the open world well

Problem:
1) we can never predict what we want in advance
2) we can never annotate all categories (especially fine-grained)

↓

Generalizing to the open world, especially for novel classes

Working pipeline:

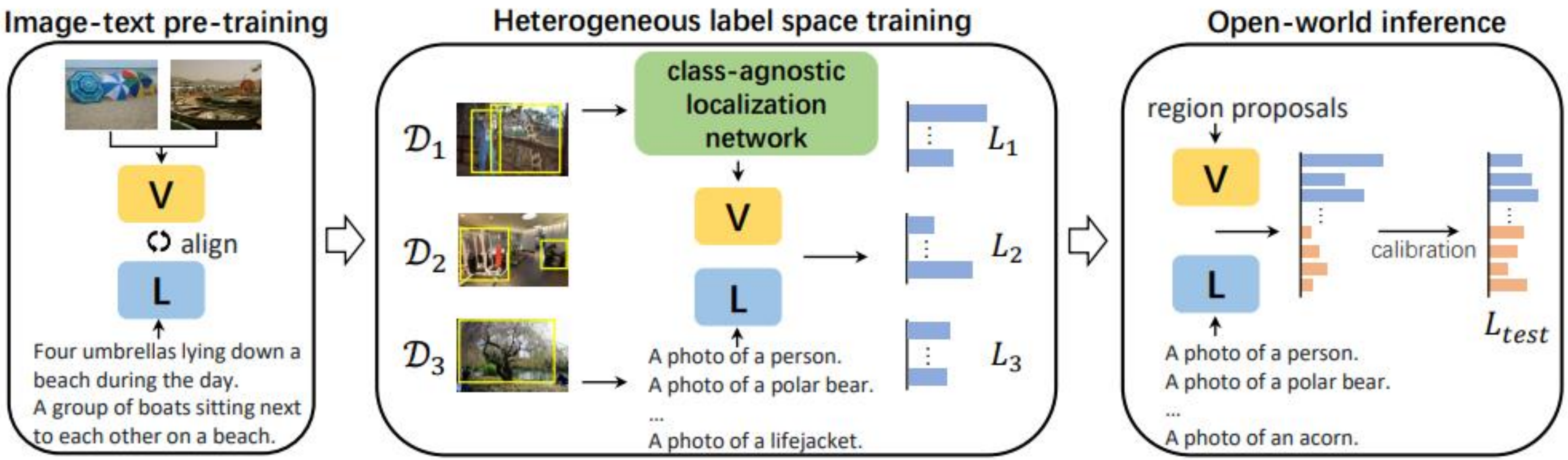


Image-text aligned pre-training

A woman with a slight smile is
picking fruits in the fruitful orchard.

An aeroplane files across the sky
In a sunny day.

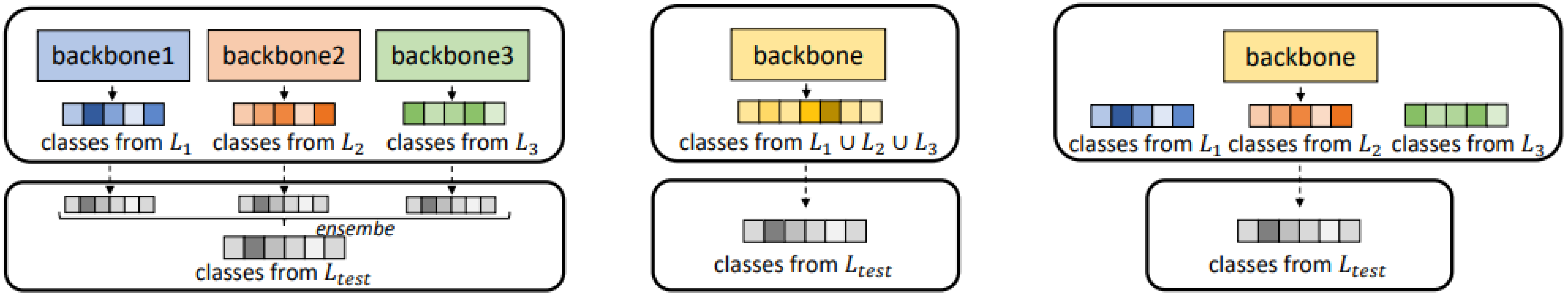
key:

image-text pairs are easy to collect (from social media)
large-scale training: see images as many as possible
align vision space and language space

CLIP, ALIGN, LiT, RegionCLIP, GLIP…

Heterogeneous label space training

Possible structures:



(a) seperate label spaces (b) unified label space (c) partitioned label space

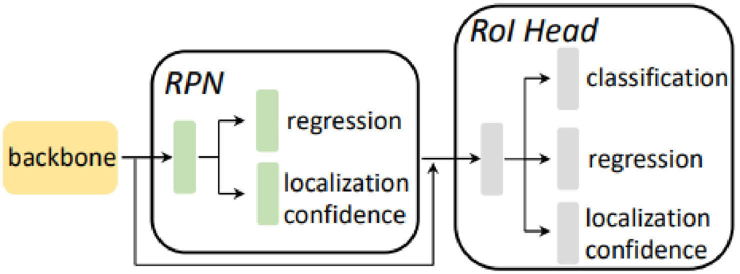
Heterogeneous label space training

Decoupling proposal generation and RoI classification:

proposal generation (RPN): class-agnostic classification
better generalize to novel classes in the open world

RoI classification: class-specific classification
cannot be generalize to the open world well

Class-agnostic localization network for proposal generation:



Open-world inference

Problem: the network is strongly biased to base classes.

Probability calibration: balance the probability prediction

$$p_{ij} = \frac{1}{1 + \exp\left(-z_{ij}^{T} e_j / \tau\right)} / \pi_j^{\gamma}$$

Experiments: open-world detection

Training datasets: subsets of COCO (80 categories), Objects365 (365 categories),
OpenImages (500 categories)
Testing datasets: LVIS v0.5 (1230 categories), v1 (1203 categories),
ImageNetBoxes (3602 categories), VisualGenome (7605 categories)

Training data	Structure	LVIS v0.5 (1,230)				LVIS v1 (1,203)				ImageNetBoxes (3,622)			VisualGenome (7,605)		
		AP	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP	AP_{50}	Loc. Acc	AR_1	AR_{10}	AR_{100}
Faster RCNN (closed world)		17.7	1.9	16.5	25.4	16.2	0.9	13.1	26.4	3.9	6.1	15.3	3.5	4.3	4.3
COCO	-	16.4	18.7	17.1	14.5	13.7	13.5	13.6	13.9	4.8	6.8	8.3	4.3	5.9	5.9
O365	-	20.2	21.3	20.2	19.8	16.8	14.7	16.2	18.3	3.8	5.5	8.4	5.4	7.3	7.3
OImg	-	16.8	21.8	17.6	13.8	13.9	14.7	14.2	13.2	7.9	10.8	16.0	5.9	8.1	8.2
COCO + O365	S	21.0	22.2	21.8	19.4	17.5	16.0	17.2	18.4	4.5	6.5	8.9	6.2	8.5	8.6
COCO + O365	U	20.9	19.6	21.0	21.3	17.6	14.6	17.0	19.6	3.6	5.1	8.0	5.3	7.1	7.2
COCO + O365 (+mosaic)	U	21.4	22.3	21.5	21.0	16.8	13.5	16.2	18.9	3.6	5.1	7.7	5.0	6.8	6.9
COCO + O365 (+pseudo [62])	U	20.8	22.5	22.7	19.7	16.6	13.4	16.1	18.7	3.6	5.1	7.6	5.0	6.6	6.7
COCO + O365	P	22.2	23.7	22.5	21.2	18.2	15.5	17.6	20.1	4.7	6.6	10.1	5.9	8.0	8.1
COCO + OImg	P	19.9	22.1	20.7	17.9	16.8	16.0	16.8	17.1	6.9	9.5	14.7	5.7	7.7	7.8
COCO + O365 + OImg	P	**23.5**	**23.6**	**24.3**	**22.4**	**19.8**	**18.0**	**19.2**	**21.2**	**8.2**	**11.4**	**16.9**	**6.5**	**8.7**	**8.8**

Experiments: closed-world detection

Training and testing on the COCO dataset:

Model	AP	AP_{50}	AP_S	AP_M	AP_L
transformer-based models					
DETR (DC5) [5]	15.5	29.4	4.3	15.1	26.7
Dynamic DETR [9]	42.9	61.0	24.6	44.9	54.4
DN-Deformable-DETR [27]	43.4	61.9	24.8	46.8	59.4
DINO [60]	49.0	66.6	32.0	52.3	63.0
CNN-based models					
Faster RCNN (FPN) [30,43]	37.9	58.8	22.4	41.1	49.1
DenseCLIP [41]	40.2	63.2	26.3	44.2	51.0
HTC [6]	42.3	61.1	23.7	45.6	56.3
Dyhead [9]	43.0	60.7	24.7	46.4	53.9
$R(Det)^2$ + Cascade [29]	42.5	61.0	24.6	45.5	57.0
Softteacher § [54]	44.5	-	-	-	-
UniDetector (ours)	**49.3**	**67.5**	**33.3**	**53.1**	**63.6**

Experiments: object detection in the wild

Inference on 13 ODinW datasets

Dataset	Objects of Interest	Train/Val/Test
PascalVOC	Common objects (PascalVOC 2012)	13690/3422/-
AerialDrone	Boats, cars, etc. from drone images	52/15/7
Aquarium	Penguins, starfish, etc. in an aquarium	448/127/63
Rabbits	Cottontail rabbits	1980/19/10
EgoHands	Hands in ego-centric images	3840/480/480
Mushrooms	Two kinds of mushrooms	41/5/5
Packages	Delivery packages	19/4/3
Raccoon	Raccoon	150/29/17
Shellfish	Shrimp, lobster, and crab	406/116/58
Vehicles	Car, bus, motorcycle, truck, and ambulance	878/250/126
Pistols	Pistol	2377/297/297
Pothole	Potholes on the road	465/133/67
Thermal	Dogs and people in thermal images	142/41/20

Model	#Data	Datasets	Avg. AP
GLIP-T (A) [28]	0.66M	Objects365	28.8
GLIP-T (B)	0.66M	Objects365	33.2
GLIP-T (C)	1.46M	Objects365, GoldG	44.4
GLIP-T	5.46M	Objects365, GoldG, Cap4M	46.5
UniDetector (ours)	**173k**	subset of COCO, Objects365, OpenImages	**47.3**

Thanks