CVPR SEATTLE, WA JUNE 17-21, 2024

Real-Time Open-Vocabulary Object Detection

Tianheng Cheng[2,3,*], Lin Song[1,2,*,✉], Yixiao Ge[1,2,🌟], Wenyu Liu[3], Xinggang Wang[3,✉], Ying Shan[1,2]

[1] Tencent AI Lab, [2] ARC Lab, Tencent PCG, [3] Huazhong University of Science and Technology

[*] Equal contribution 🌟 Project lead ✉ Corresponding author

thch@hust.edu.cn

ARC Applied Research Center

Tencent AI Lab

华中科技大学 HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Traditional Object Detection

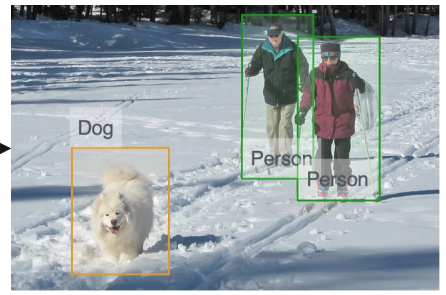
Traditional detectors are capable to detect objects in **close-set datasets**

Images or Videos

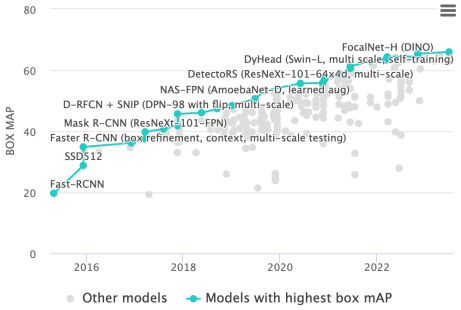
Object Detector

Faster R-CNN, YOLO, CenterNet,

FCOS, DETR, Deformable DETR

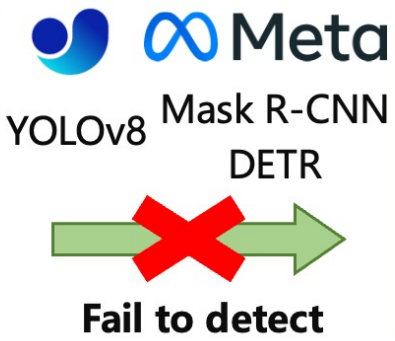


Classification & Localization



66.0 COCO AP

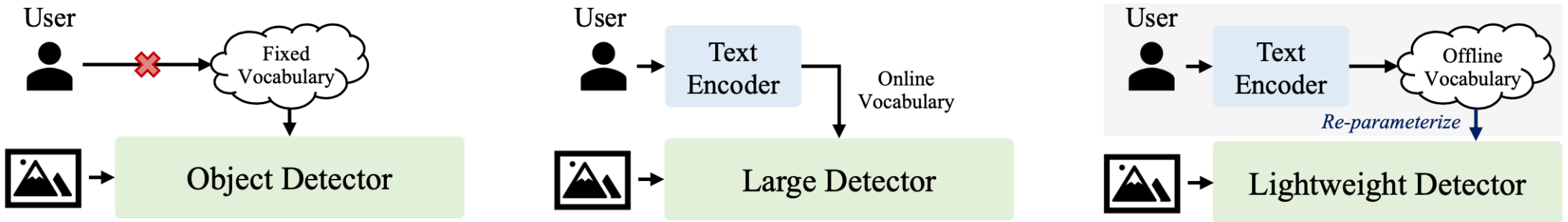
Traditional detectors **fail** to detect objects **not appeared** in the training data



T. Cheng et al. YOLO-World: Real-Time Open-Vocabulary Object Detection. CVPR 2024.

Open-Vocabulary Object Detection

➢ **Efficient *Prompt-then-Detect* Paradigm**



(a) Traditional Object Detector (b) Preivous Open-Vocabulary Detector (c) YOLO-World

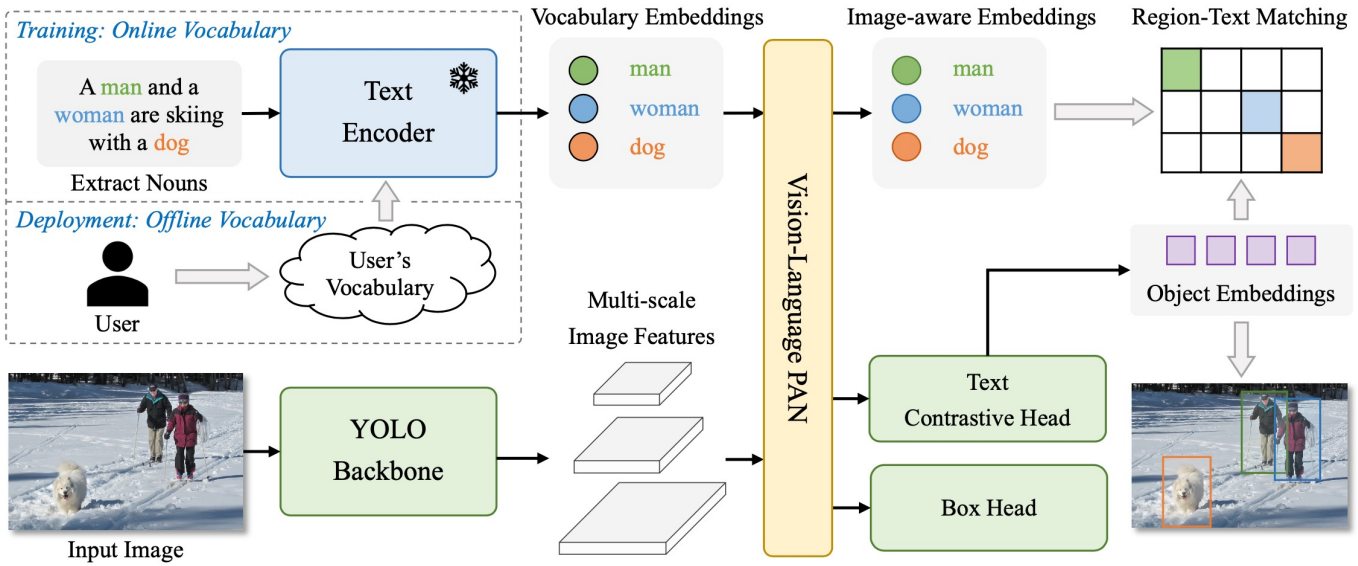
Traditional Detectors: fixed vocabulary, user can not modify.

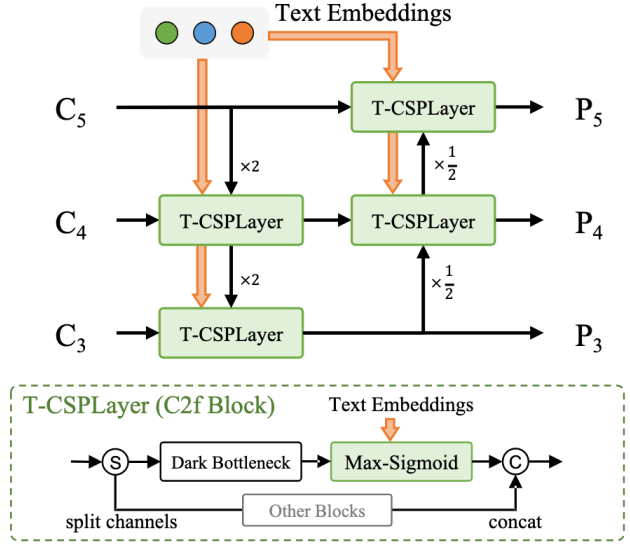
Previous Open-Vocabulary Detectors: detectors are heavy and in-efficient, detect based on users' prompts (forward text encoders).

YOLO-World: aim for efficiency, user can modify the vocabulary on demand.

YOLO-World

➢ Model Architecture

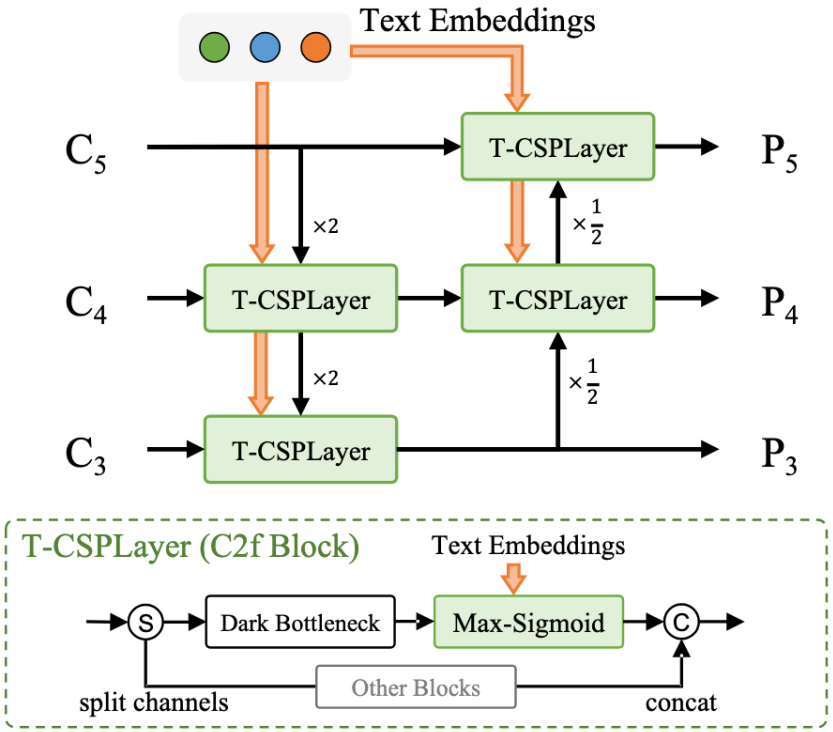




- **Open-Vocabulary YOLO detector** with a text encoder and vision-language modeling
- A **re-parameterizable vision-language feature pyramid networks (RepVL-PAN)** for fusing language information into image features.
- Pre-training on large-scale **region-text pairs**: detection, grounding, and image-text data.

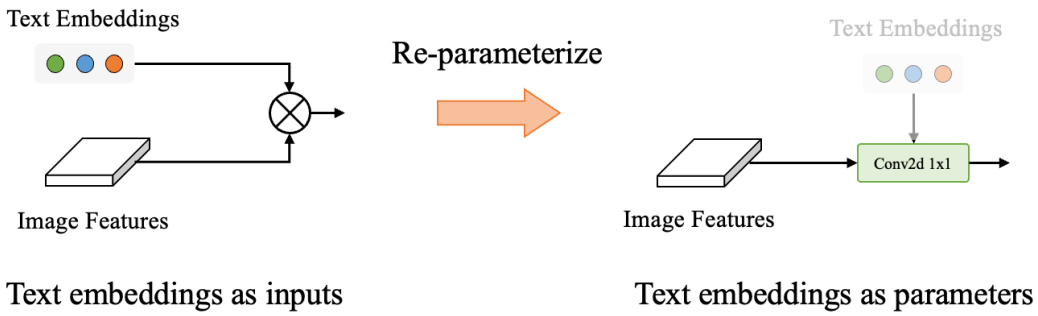
YOLO-World

➢ Model Architecture: RepVL-PAN



RepVL-PAN

- **Text-guided Layer** to fuse text embeddings into image features.
- Efficient deployment through **re-parameterization**.

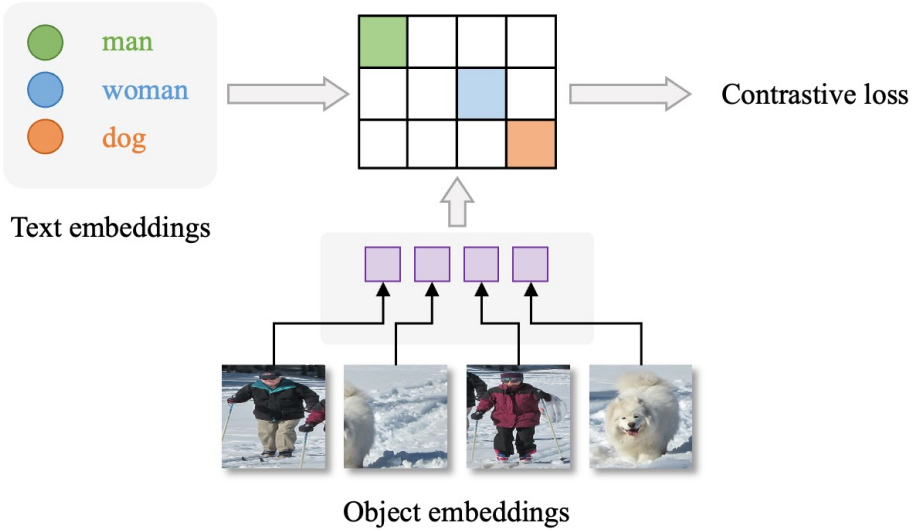


re-parameterization: dot-product as a simple conv

YOLO-World

➢ Region-Text Training

Contrastive Loss on Region-Text Pairs



Pre-training Datasets: nearly 1.6M samples

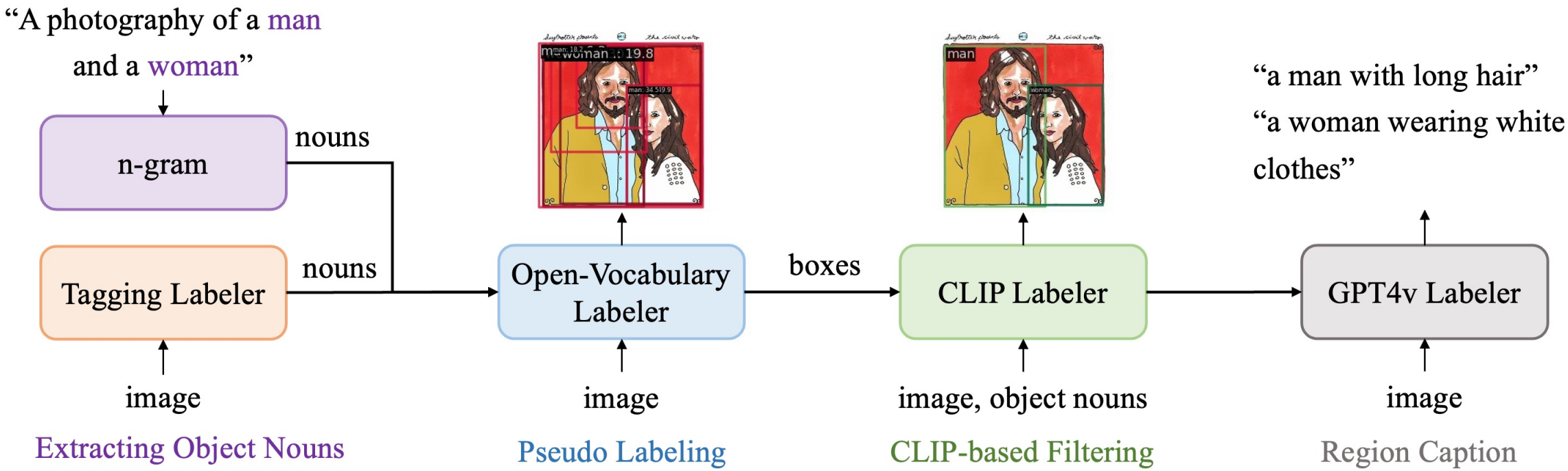
Dataset	Type	Vocab.	Images	Anno.
Objects365V1 [46]	Detection	365	609k	9,621k
GQA [17]	Grounding	-	621k	3,681k
Flickr [38]	Grounding	-	149k	641k
CC3M† [47]	Image-Text	-	246k	821k

Auto Labeling

YOLO-World

➢ Data Engine

Automatic labeling large-scale image-text pairs: **region-text (caption) pairs**

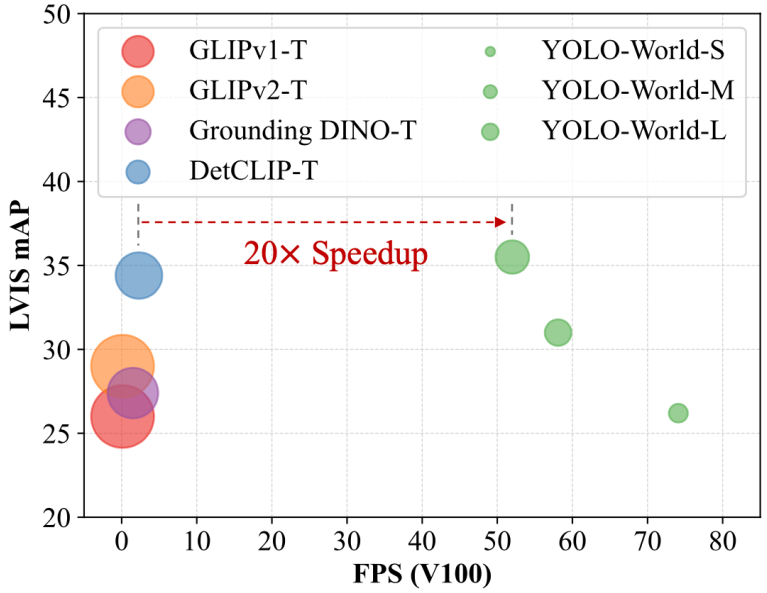


YOLO-World

➢ Experimental Results: Zero-Shot Object Detection

[1] Zero-shot Evaluation on LVIS (1203 categories)

Method	Backbone	Params	Pre-trained Data	FPS	AP	AP_r	AP_c	AP_f
MDETR [19]	R-101 [14]	169M	GoldG	-	24.2	20.9	24.3	24.2
GLIP-T [22]	Swin-T [30]	232M	O365,GoldG	0.12	24.9	17.7	19.5	31.0
GLIP-T [22]	Swin-T [30]	232M	O365,GoldG,Cap4M	0.12	26.0	20.8	21.4	31.0
GLIPv2-T [54]	Swin-T [30]	232M	O365,GoldG	0.12	26.9	-	-	-
GLIPv2-T [54]	Swin-T [30]	232M	O365,GoldG,Cap4M	0.12	29.0	-	-	-
Grounding DINO-T [28]	Swin-T [30]	172M	O365,GoldG	1.5	25.6	14.4	19.6	32.2
Grounding DINO-T [28]	Swin-T [30]	172M	O365,GoldG,Cap4M	1.5	27.4	18.1	23.3	32.7
DetCLIP-T [51]	Swin-T [30]	155M	O365,GoldG	2.3	34.4	26.9	33.9	36.3
YOLO-World-S	YOLOv8-S	13M (77M)	O365,GoldG	74.1 (19.9)	26.2	19.1	23.6	29.8
YOLO-World-M	YOLOv8-M	29M (92M)	O365,GoldG	58.1 (18.5)	31.0	23.8	29.2	33.9
YOLO-World-L	YOLOv8-L	48M (110M)	O365,GoldG	52.0 (17.6)	35.0	27.1	32.8	38.3
YOLO-World-L	YOLOv8-L	48M (110M)	O365,GoldG,CC3M†	52.0 (17.6)	35.4	27.6	34.1	38.0



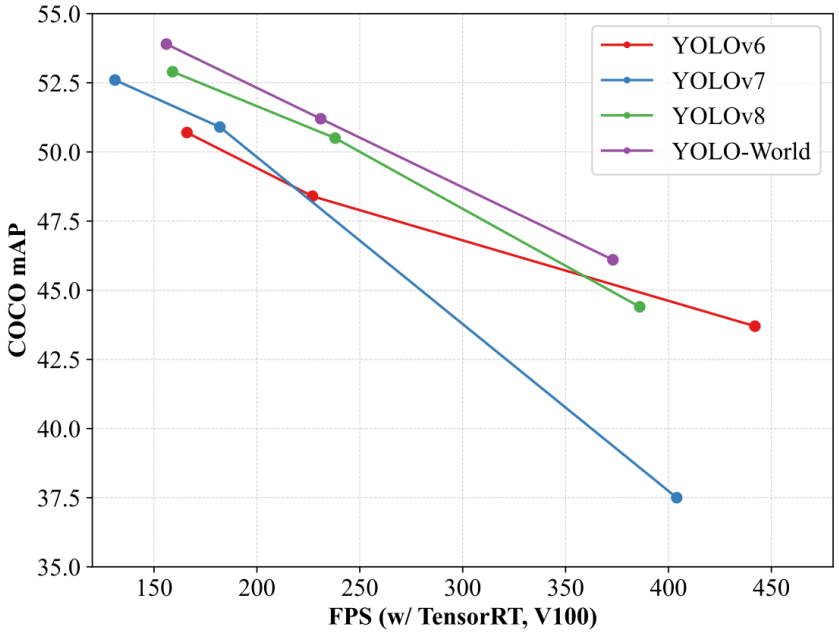
YOLO-World obtains **comparable zero-shot accuracy** with heavy detectors while achieving **remarkable inference speed**!

YOLO-World

➢ Experimental Results: Zero-Shot/Finetuned Object Detection

[2] COCO Zero-shot & Fine-tuning

Model	Epochs	AP^{zero}	AP	AP_{50}	AP_{75}
YOLOv8-S	500	-	44.4	61.2	48.1
YOLO-World-S	80	37.5	**46.1**	62.0	49.9
YOLOv8-M	500	-	50.5	67.3	55.0
YOLO-World-M	80	42.8	**51.0**	67.5	55.2
YOLOv8-L	500	-	52.9	69.9	67.7
YOLO-World-L	80	45.4	**53.9**	70.9	58.8



YOLO-World obtains competitive zero-shot performance on COCO, and significantly **better fine-tuned performance** compared to YOLOv8 (baseline) or other YOLO detectors.

YOLO-World

➢ Experimental Results: Ablations

[3] Data Scaling

Pre-trained Data	AP	AP_r	AP_c	AP_f
O365	23.5	16.2	21.1	27.0
O365,GQA	31.9	22.5	29.9	35.4
O365,GoldG	32.5	22.3	30.6	**36.0**
O365,GoldG,CC3M†	**33.0**	**23.6**	**32.0**	35.5

- Pre-training with more **rich texts** (GQA / GoldG) improves open-vocabulary ability (AP_r: rare AP).

[4] Text-guidance from RepVL-PAN

Data	Text-guided?	AP	AP_r	AP_c	AP_f
O365	✗	22.4	14.5	20.1	26.0
O365	✓	23.2	15.2	20.6	27.0
O365+GG	✗	30.9	19.8	29.1	34.6
O365+GG	✓	32.6	27.8	31.1	34.9

- Adding **text-guidance** contributes to detecting novel objects (AP_r: rare AP), especially works well with datasets containing rich texts (GoldG).

YOLO-World

➢ Experimental Results: Zero-Shot Grounding

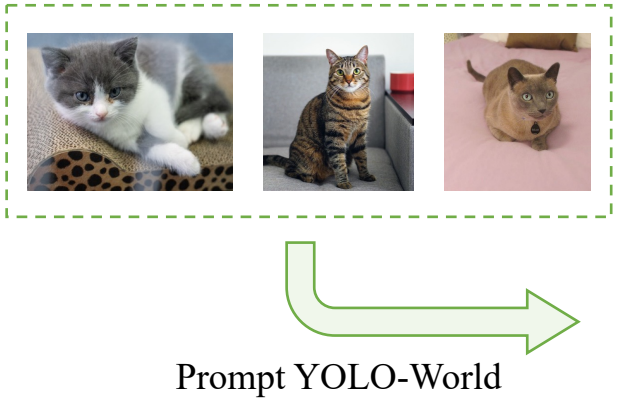
[5] Vision Grounding

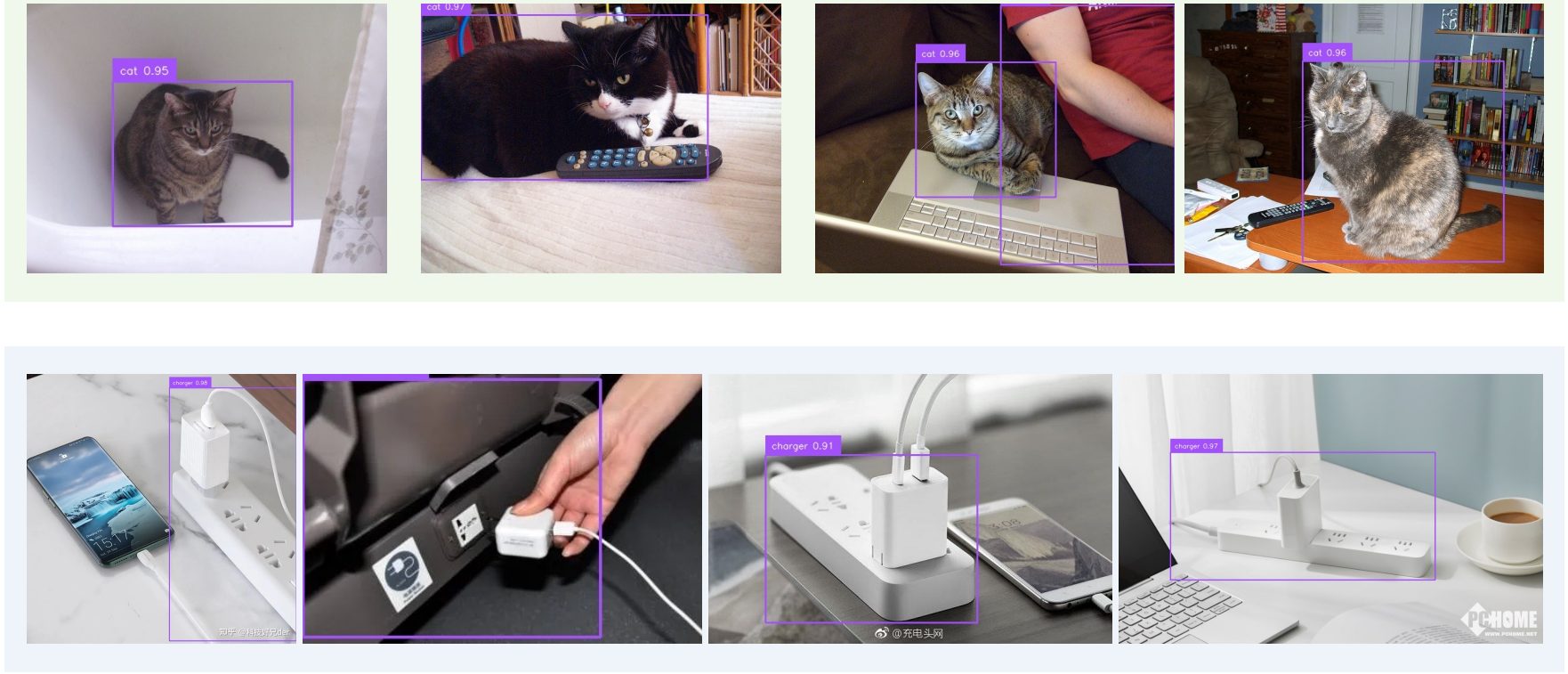
Zero-shot **grounding ability** and **language understanding** ability

YOLO-World

➢ YOLO-World as Few-shot Learners

YOLO-World supports **few image prompts**, detect objects based on referenced images





T. Cheng et al. YOLO-World: Real-Time Open-Vocabulary Object Detection. CVPR 2024.

YOLO-World

➢ **YOLO-World-SAM**

“mammoth”

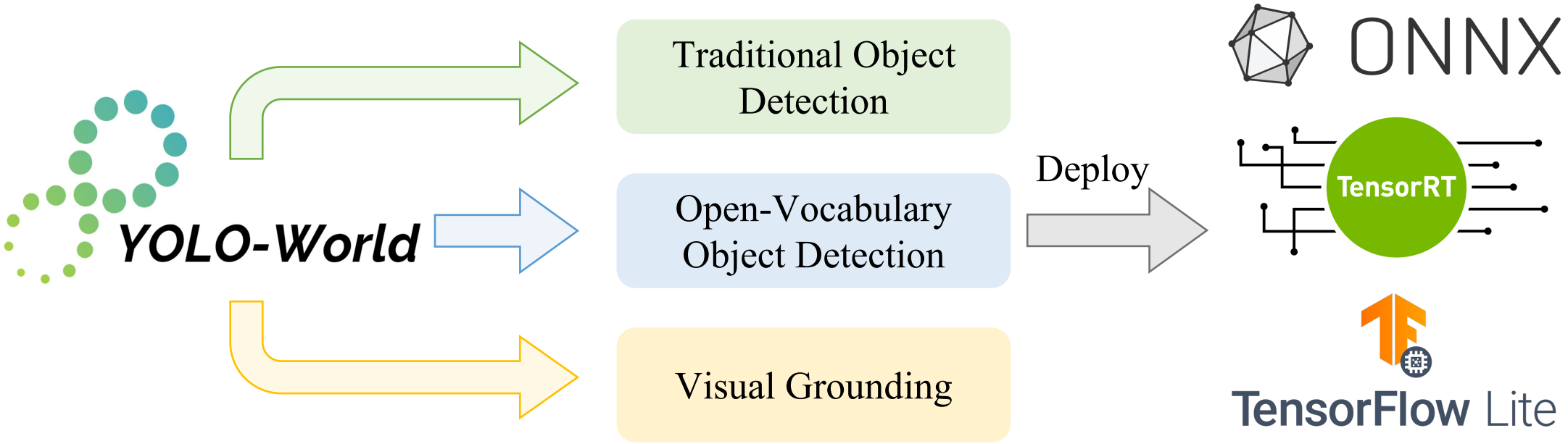
“the bigger cat”

YOLO-World + SAM + SD

Many interesting projects are developed based on YOLO-World!

YOLO-World

➢ **Boost Real-World Applications**



One YOLO-World, More Applications

QR Codes

Homepage

arXiv paper

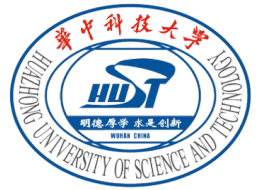
Code & Models

🤗 Demo

Thanks


ARC
Applied
Research
Center


Tencent
AI Lab


HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY