

Scaling Open-Vocabulary Object Detection

Authors: Matthias Minderer, Alexey Gritsenko, and Neil Houlsby

NeurIPS 2023, 23 citations

Presented by Group 7: Daniel Cisneros, Suranadi Dodampagamage, Andrew El-Kommos, Bradley Racey, Salem Long



Outline

1. Background
2. Method
3. Results
4. Limitations
5. Conclusion



Background



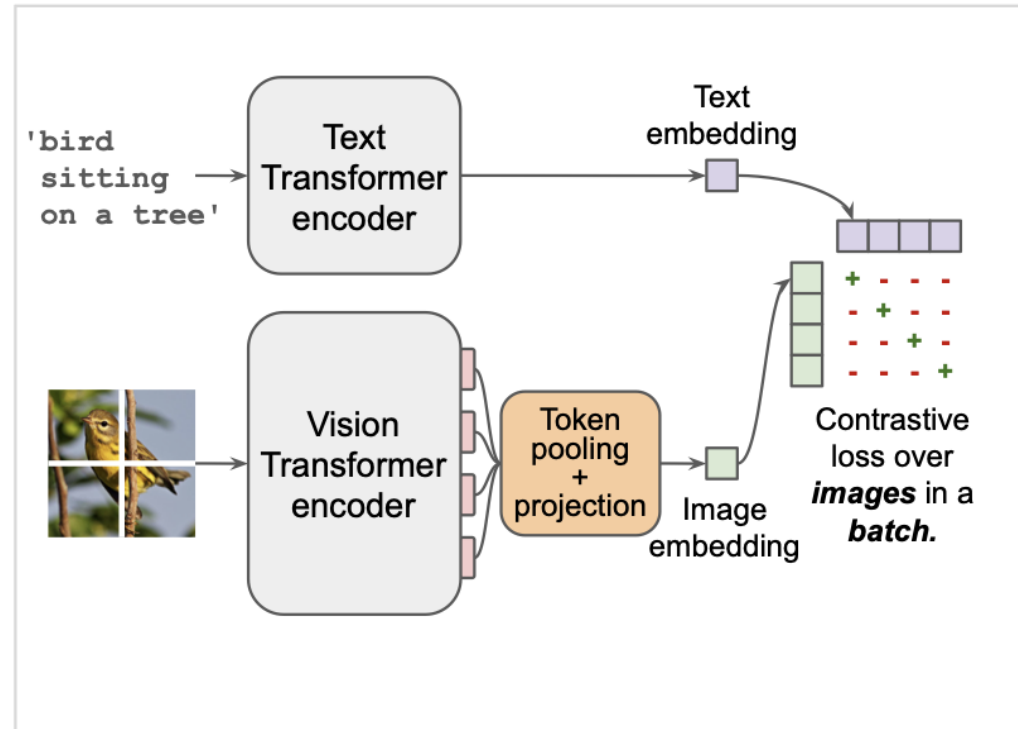
Open Vocabulary Detection



Cake with cherries on top

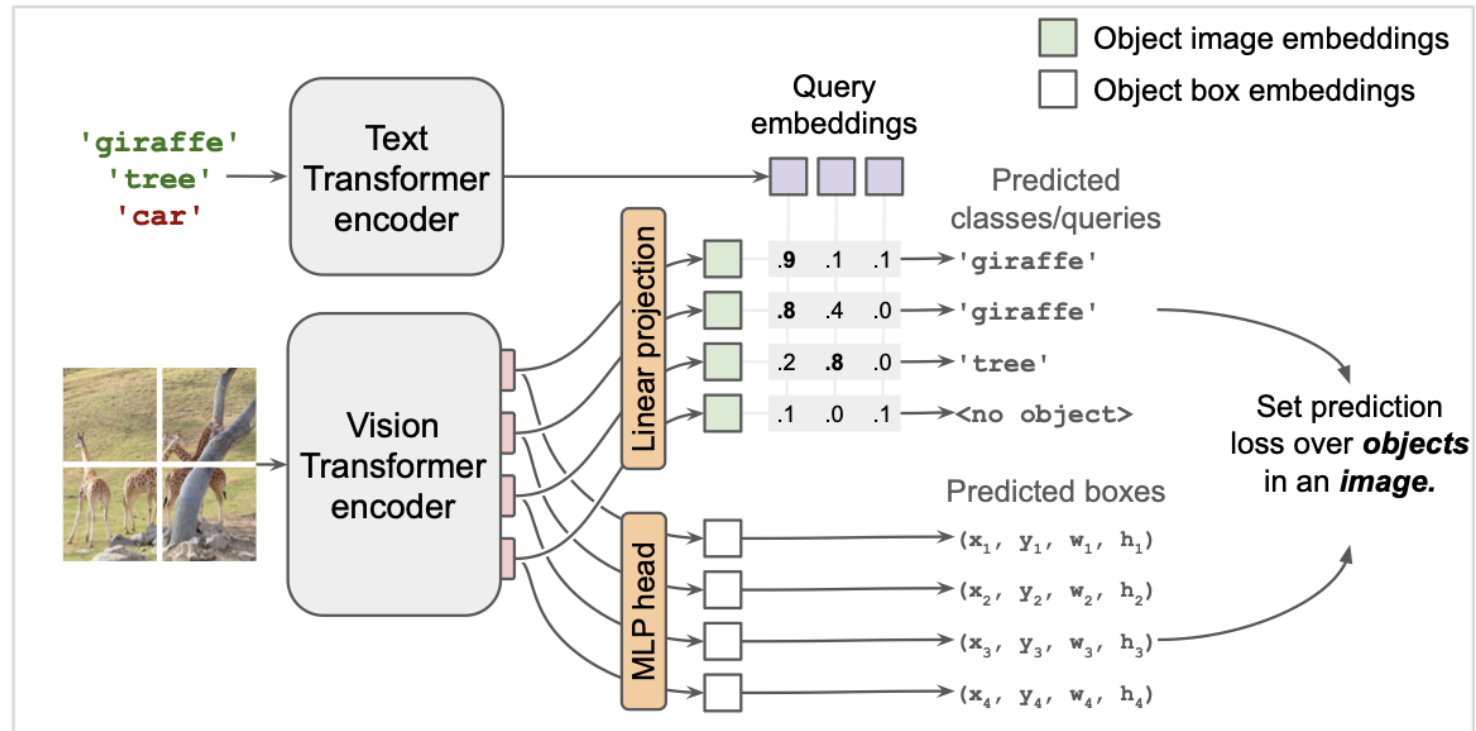
OWL-ViT Architecture

Image-level contrastive pre-training



OWL-ViT Architecture

Transfer to open-vocabulary detection



Self-Training, Label Spaces, and N-Grams

- How can the scarcity of detection data can be addressed?
- Self-training, uses an existing detector to predict bounding boxes on unlabeled images to generate data for training better detectors.
- Label spaces, refers to the set of all possible labels that can be assigned to instances in a dataset.



Self-Training, Label Spaces, and N-Grams

- Expanded label spaces using N-Grams



Unigram: Apple

Bigram: Red apple

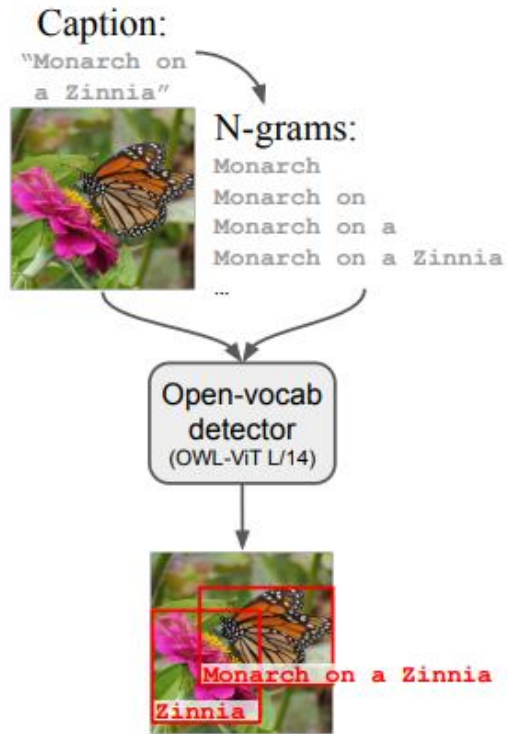
Trigram: Juicy red apple

Method



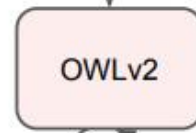
Method

1. Annotation



2. Self-training

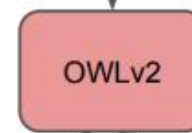
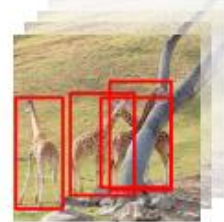
Pseudo-annotated
WebLI



OWL-ViT
detection loss

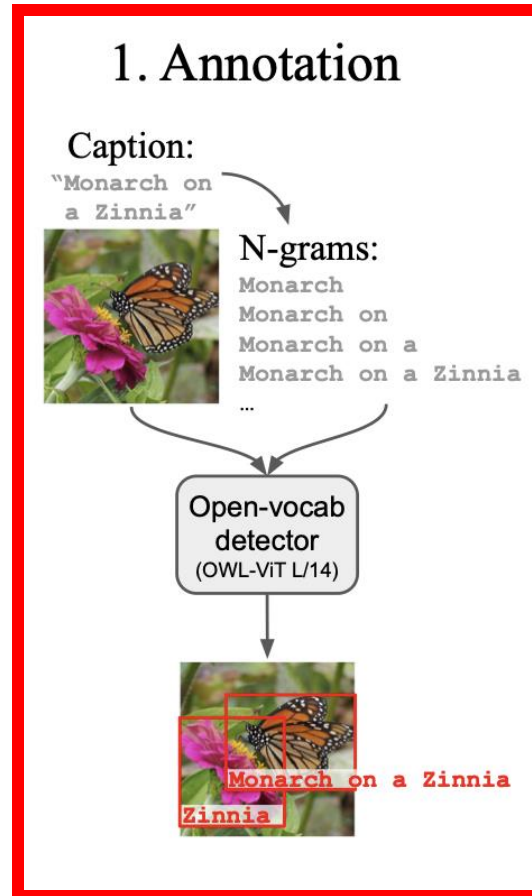
3. Fine-tuning *(optional)*

LVIS_{base}



OWL-ViT
detection loss

Annotation



2. Self-training

Pseudo-annotated
WebLI

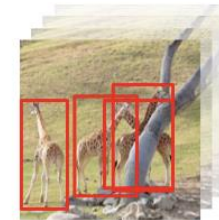


OWLv2

OWL-ViT
detection loss

3. Fine-tuning (optional)

LVIS_{base}



OWLv2

OWL-ViT
detection loss

Annotations

Dataset :

- WebLI (Image-Text Dataset)
- ~10 B images

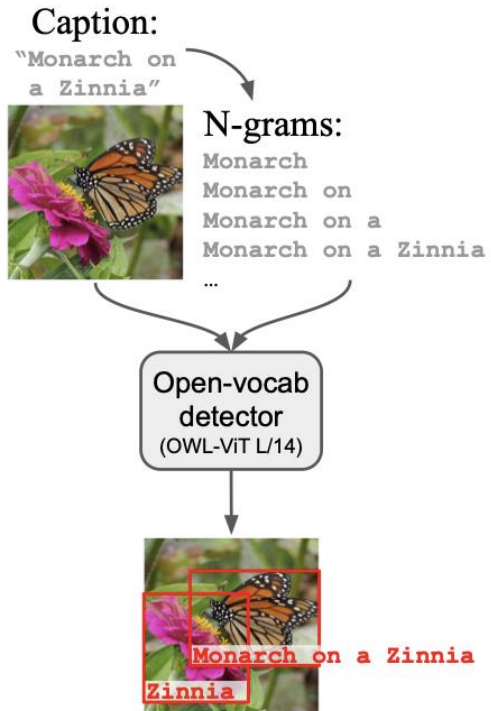
Label Spaces :

- Human-curated label space
- Machine-generated label space



Self-training

1. Annotation



2. Self-training

Pseudo-annotated
WebLI

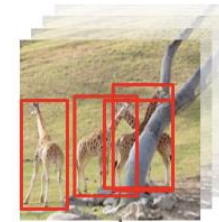


OWLv2

OWL-ViT
detection loss

3. Fine-tuning (optional)

LVIS_{base}



OWLv2

OWL-ViT
detection loss

Self-training

- Self-train a new detector on the pseudo-annotations
- Enhance training efficiency with :
 - Token dropping
 - Instance selection
 - Mosaics



Token dropping

- Images contains low pixel variance areas
- Less informative
- Drop 50% tokens lower than mean pixel variance



Token dropping

Metric	Token drop rate				
	0.00	0.25	0.33	0.50	0.70
LVIS AP _{all} ^{val}	33.3 ±0.33	33.1	33.6	32.9	30.4
LVIS AP _{rare} ^{val}	31.8 ±1.16	31.0	32.6	30.8	28.2

- Dropping up to 50% of tokens does not significantly affect performance
- It remains within one standard deviation of the full performance

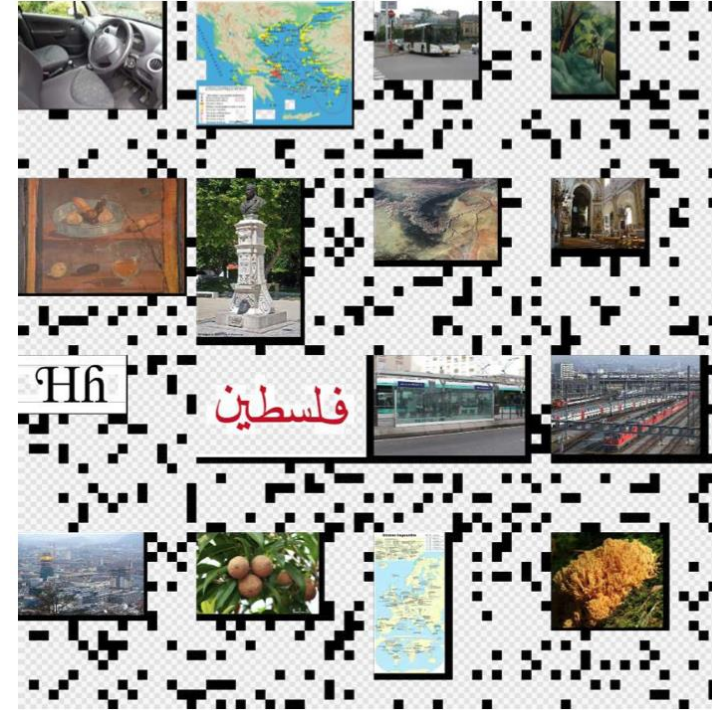


Instance selection

- OWL-ViT predicts one bounding box per encoder token
- Most output tokens do not represent objects
- Objectness head: predicts likelihood that an output token represents an object
- Select 10% of instances by top objectness during training



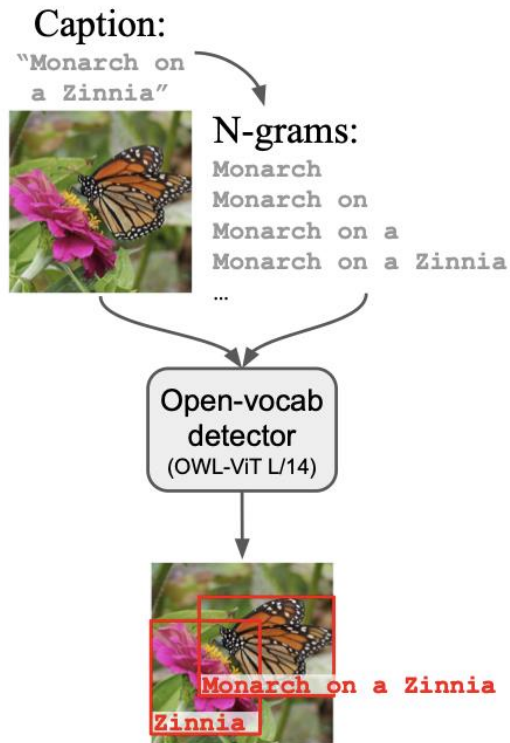
Mosaics



4 x 4 mosaic before and after 50% of patches dropped

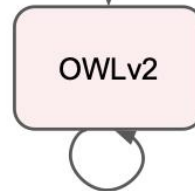
Fine-tuning

1. Annotation



2. Self-training

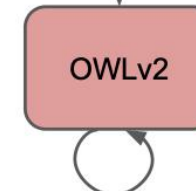
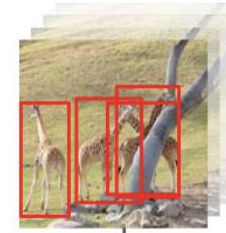
Pseudo-annotated
WebLI



OWL-ViT
detection loss

3. Fine-tuning (optional)

LVIS_{base}



OWL-ViT
detection loss

Fine-Tuning

- Further improves detection performance.
- Trade-off between open vocabulary performance and fine-tuned classes.
- Create an ensemble of the model by averaging model weights.



Results



Experiments – results

Method	Backbone	Self-training data	Self-training vocabulary	Human box annotations	ODinW 13	LVIS AP _{all} ^{mini}	LVIS AP _{rare} ^{mini}	LVIS AP _{all} ^{val}	LVIS AP _{rare} ^{val}	
<i>Open vocabulary (evaluation vocabulary is not available at training time):</i>										
1	RegionCLIP [40]	R50x4	CC3M	6k concepts	LVIS _{base}	–	–	–	32.3	22.0
2	OWL [21]	CLIP B/16	–	–	O365+VG	–	–	–	27.2	20.6
3	OWL [21]	CLIP L/14	–	–	O365+VG	48.4	–	–	34.6	31.2
4	GLIPv2 [39]	Swin-T	Cap4M	tokens	O365+GoldG	48.5	29.0	–	–	–
5	GLIPv2 [39]	Swin-B	CC15M	tokens	FiveODs+GoldG	54.2	48.5	–	–	–
6	GLIPv2 [39]	Swin-H	CC15M	tokens	FiveODs+GoldG	55.5	50.1	–	–	–
7	F-VLM [14]	R50x4	–	–	LVIS _{base}	–	–	–	28.5	26.3
8	F-VLM [14]	R50x64	–	–	LVIS _{base}	–	–	–	34.9	32.8
9	3Ways [1]	NFNet-F0	TODO	captions	LVIS _{base}	–	–	–	35.7	25.6
10	3Ways [1]	NFNet-F6	TODO	captions	LVIS _{base}	–	–	–	44.6	30.1
11	OWL-ST	CLIP B/16	WebLI	N-grams	O+VG	48.8	31.8	35.4	27.0	29.6 -3.2
12	OWL-ST	CLIP L/14	WebLI	N-grams	O+VG	53.0	38.1	39.0	33.5	34.9 +2.1
13	OWL-ST	SigLIP G/14	WebLI	N-grams	O+VG	49.9	37.8	40.9	33.7	37.5 +4.7

9 point improvement for OWL-ST compared with previous OWL-ViT even without fine-tuning!



Experiments – results

Method	Backbone	Self-training data	Self-training vocabulary	Human box annotations	ODinW 13	LVIS AP ^{mini} _{all}	LVIS AP ^{mini} _{rare}	LVIS AP ^{val} _{all}	LVIS AP ^{val} _{rare}		
11	OWL-ST	CLIP B/16	WebLI	N-grams	O+VG	48.8	31.8	35.4	27.0	29.6	+3.2
12	OWL-ST	CLIP L/14	WebLI	N-grams	O+VG	53.0	38.1	39.0	33.5	34.9	+2.1
13	OWL-ST	SigLIP G/14	WebLI	N-grams	O+VG	49.9	37.8	40.9	33.7	37.5	+4.7
14	OWL-ST+FT	CLIP B/16	WebLI	N-grams	O+VG, LVIS _{base}	48.6	47.2	37.8	41.8	36.2	+3.4
15	OWL-ST+FT	CLIP L/14	WebLI	N-grams	O+VG, LVIS _{base}	50.1	54.1	46.1	49.4	44.6	+11.8
16	OWL-ST+FT	SigLIP G/14	WebLI	N-grams	O+VG, LVIS _{base}	50.1	51.3	50.9	47.0	47.2	+14.4

- Why does the model perform better on LVIS_{rare} than on LVIS_{all}?



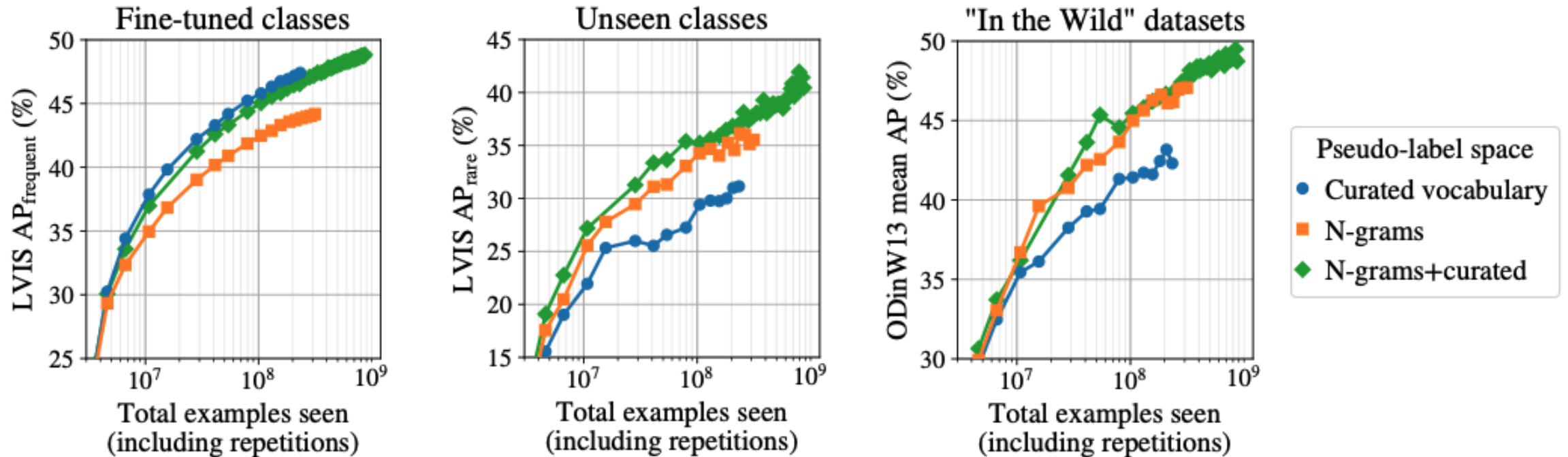
Experiments – results

Method	Backbone	Self-training data	Self-training vocabulary	Human box annotations	ODinW 13	LVIS AP ^{mini} _{all}	LVIS AP ^{mini} _{rare}	LVIS AP ^{val} _{all}	LVIS AP ^{val} _{rare}
7 F-VLM [14]	R50x4	–	–	LVIS _{base}	–	–	–	28.5	26.3
8 F-VLM [14]	R50x64	–	–	LVIS _{base}	–	–	–	34.9	32.8
9 3Ways [1]	NFNet-F0	TODO	captions	LVIS _{base}	–	–	–	35.7	25.6
10 3Ways [1]	NFNet-F6	TODO	captions	LVIS _{base}	–	–	–	44.6	30.1
11 OWL-ST	CLIP B/16	WebLI	N-grams	O+VG	48.8	31.8	35.4	27.0	29.6 -3.2
12 OWL-ST	CLIP L/14	WebLI	N-grams	O+VG	53.0	38.1	39.0	33.5	34.9 +2.1
13 OWL-ST	SigLIP G/14	WebLI	N-grams	O+VG	49.9	37.8	40.9	33.7	37.5 +4.7
14 OWL-ST+FT	CLIP B/16	WebLI	N-grams	O+VG, LVIS _{base}	48.6	47.2	37.8	41.8	36.2 +3.4
15 OWL-ST+FT	CLIP L/14	WebLI	N-grams	O+VG, LVIS _{base}	50.1	54.1	46.1	49.4	44.6 +11.8
16 OWL-ST+FT	SigLIP G/14	WebLI	N-grams	O+VG, LVIS _{base}	50.1	51.3	50.9	47.0	47.2 +14.4

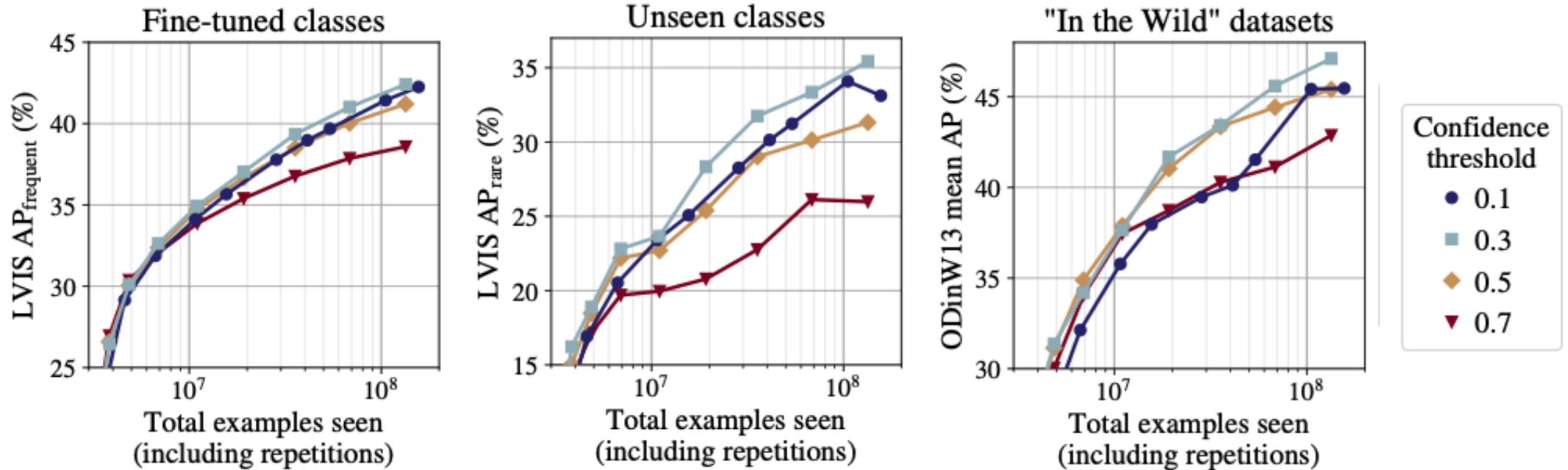
With fine-tuning, OWL-ST is 14.4 points higher than the next best model from literature!



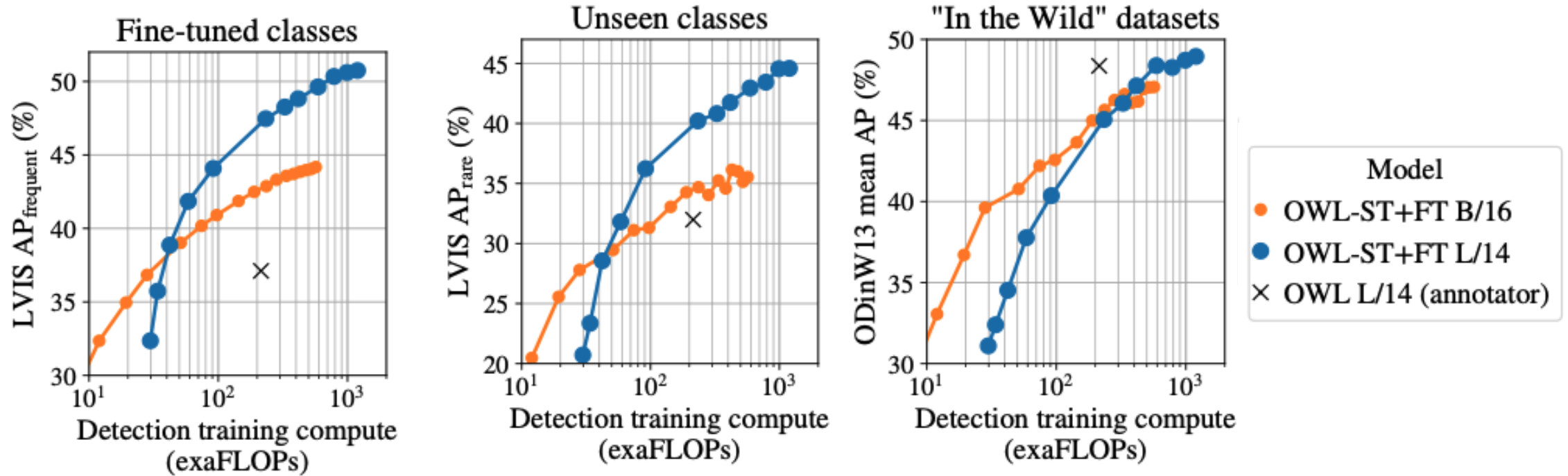
Experiments – pseudo-annotation label space



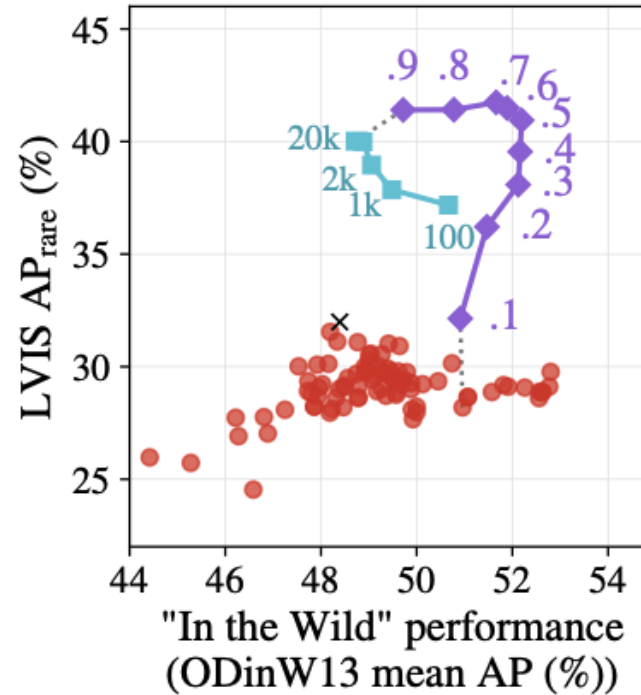
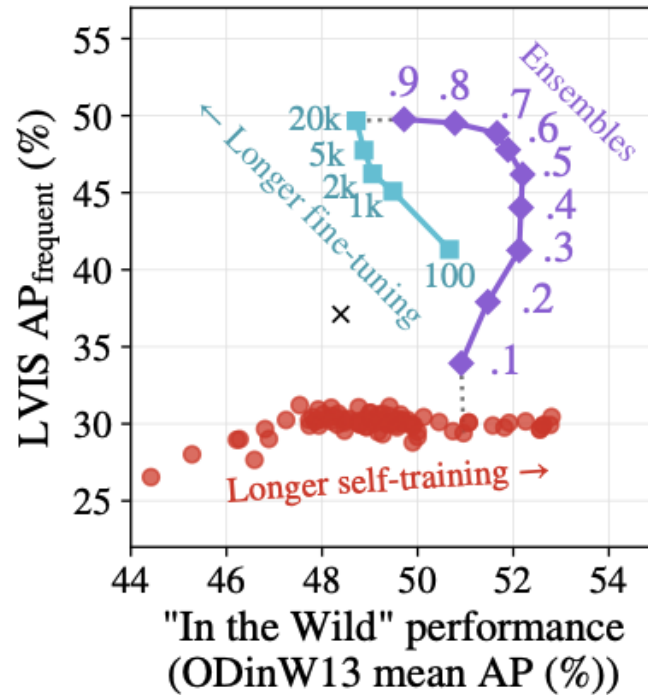
Experiments – pseudo-annotations filtering



Experiments – scaling



Experiments – effects of fine-tuning



- Self-training only
- Fine-tuned on LVIS base
- ◆ Weight ensemble
- × OWL L/14 (annotator)

Experiments – effects of fine-tuning

OWL-ST L/14 self-trained on N-grams, not fine-tuned (Table 1 row 12)



OWL-ST+FT L/14 self-trained on N-grams and fine-tuned on LVIS_{base} (Table 1 row 15)



Experiments – effects of fine-tuning

OWL-ST L/14 self-trained on N-grams, not fine-tuned (Table 1 row 12)

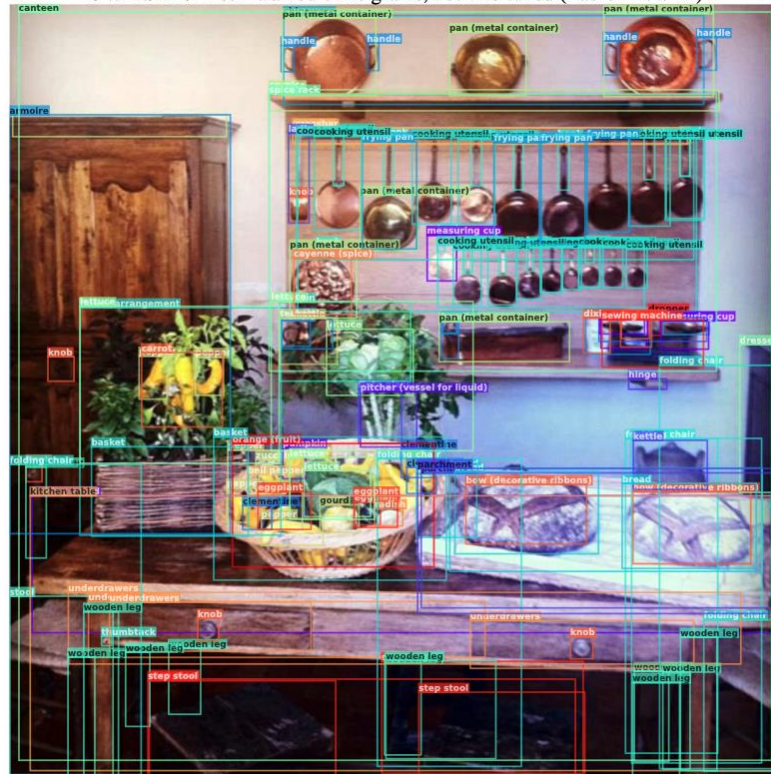


OWL-ST+FT L/14 self-trained on N-grams and fine-tuned on LVIS_{base} (Table 1 row 15)



Experiments – effects of fine-tuning

OWL-ST L/14 self-trained on N-grams, not fine-tuned (Table 1 row 12)



OWL-ST+FT L/14 self-trained on N-grams and fine-tuned on LVIS_{base} (Table 1 row 15)



Limitations

- Massive amount of computational resources and data required for self-training
- Trade-off between fine-tuned and open-vocabulary performance (discussed previously)



Conclusion

- Self-training can be scaled up to overcome dependency on human annotations
- OWL-ST shows significant improvements in detection performance using weak supervision from web data



Thank you!

