Open-Vocabulary Detection & Segmentation

OWL-ViT – *NeurIPS '22*

LSeg – *ICLR '22*

DetCLIP-v3 – *CVPR '24*

Presenters: Alexander Karpekov, Kasra Sohrab, Kausar Patherya

CS8803: VLM

Fall 2025

Agenda

Agenda

- Introductions
- Background & Definitions
  - Timeline
  - Classification vs Detection vs Segmentation
  - Metrics & Datasets
- Papers Deep Dive:
  - OWL-ViT
  - LSeg
  - DetCLIP-v3
- Summary

Presenters

Alexander Karpekov

- **CS PhD** [website]
- ***Advisors***: Thomas Plötz & Sonia Chernova
- ***Interests***: Explainable AI, Computational Theory of Mind, Visualizations

Kasra Sohrab

- **CS Masters** (Thesis) [LinkedIn]
- ***Advisor***: Alexey Tumanov
- ***Interest***: Systems for AI, currently LLMs

Kausar Patherya

- **Robotics/CS Masters** [website]
- ***Advisor***: Lu Gan & Matthew Gombolay
- ***Interests***: Semantic Mapping, Causal RL, Embedded AI

Background & Definitions

How would you describe this image?

Source: Instagram

VLM Timeline: What have we seen so far?

VLM Backbones: Classification

ViT
flat global self-attention for images

Swin T
hierarchical Transformer With shifter windows

Generative Multimodal: Create Captions

BLIP-1
unifying pre-training for captioning + retrieval + VQA

BLIP-2
Q-Former, frozen image encoders with LLMs

Flamingo
few-shot multimodal LLM

BLIP-3
Stronger LLM

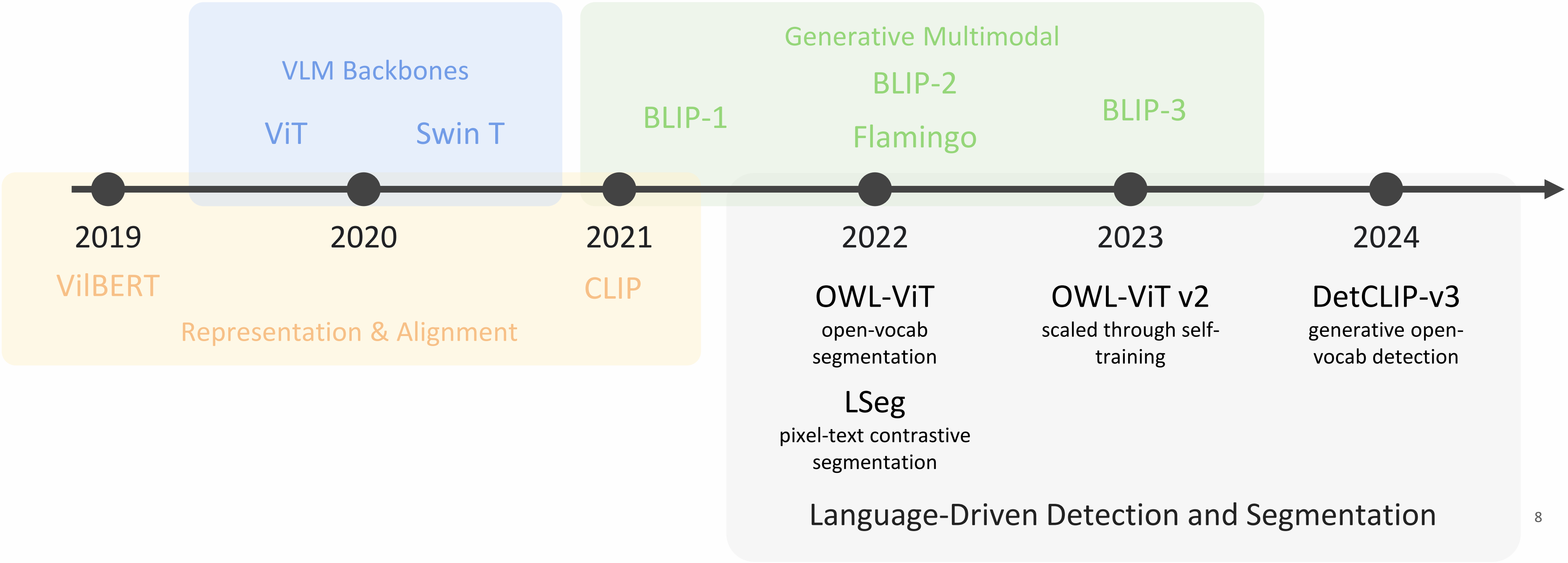
2019 — 2020 — 2021 — 2022 — 2023 — 2024

ViLBERT
early two-stream fusion model

CLIP
contrastive image–text embeddings at scale

Representation & Alignment

VLM Timeline: What will we cover today?

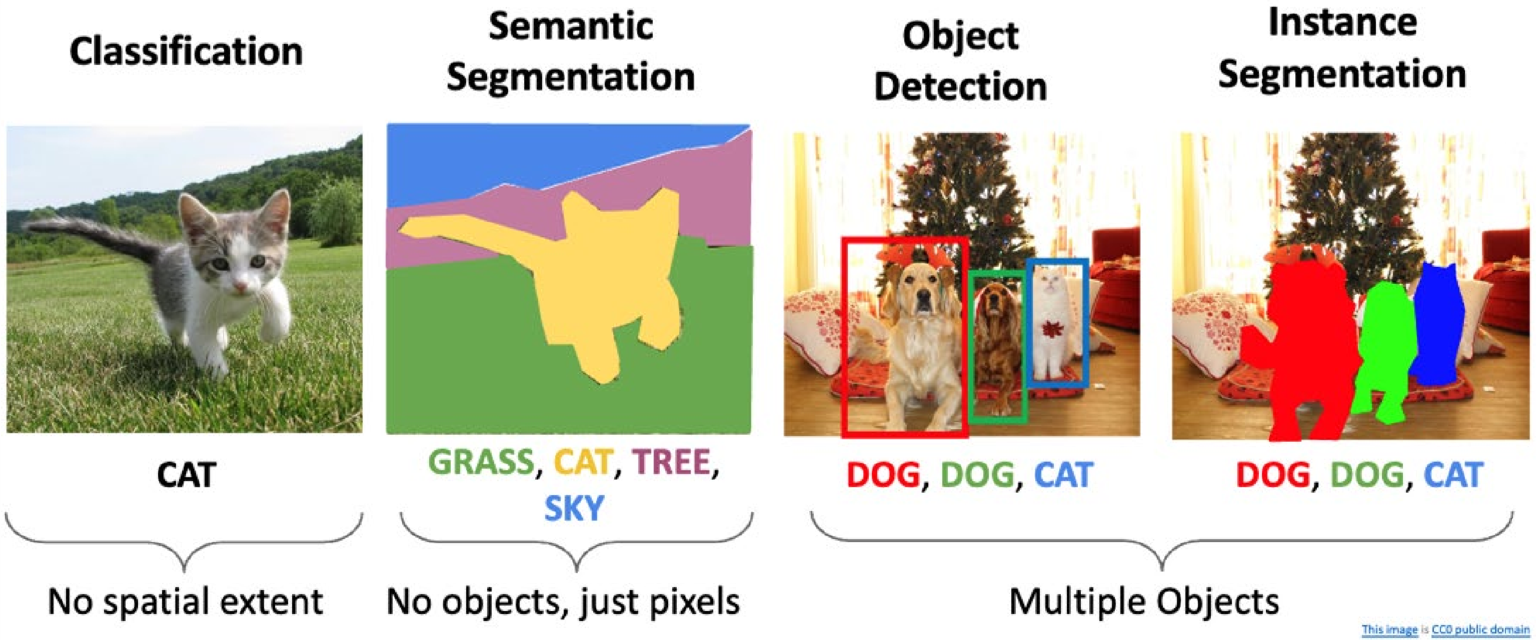


Computer Vision Tasks

OWL-ViT

LSeg

DetCLIP-v3

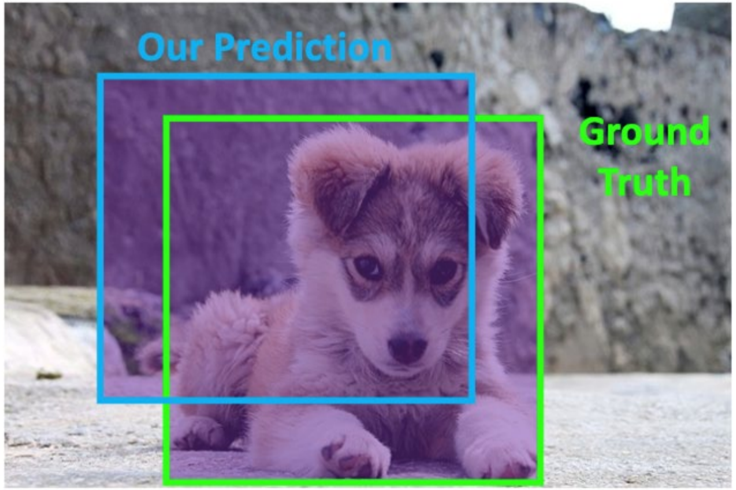


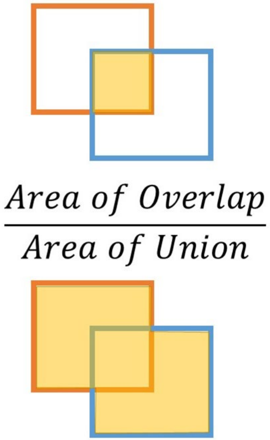
Source: UMich CV Lecture Slides

Metrics: IoU and AP50

(Average Precision at 50% Intersection over Union)

IoU: Intersection over Union





AP: Average Precision

Metrics	Metrics Meaning
AP	AP at IoU = 0.50: 0.05: 0.95
AP_{50}	AP at IoU = 0.50
AP_{75}	AP at IoU = 0.75
AP_S	AP for small objects: area < 3

Source: UMich CV Lecture Slides

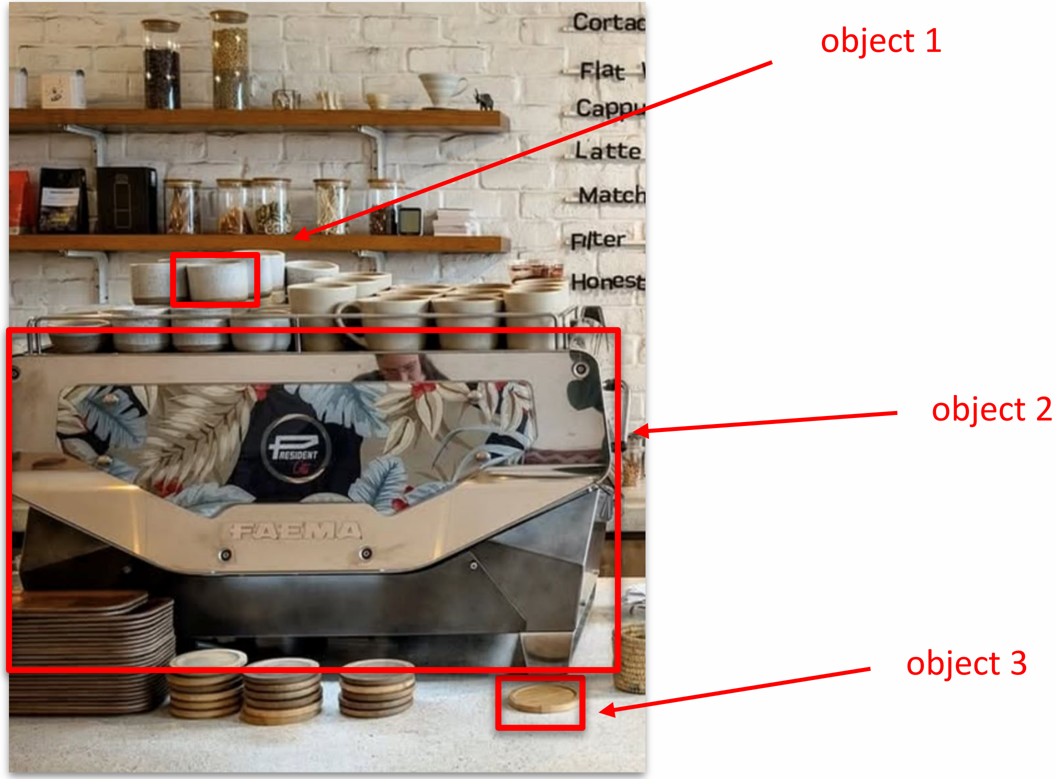
OWL-ViT

Simple Open-Vocabulary Object Detection with Vision Transformers

What objects do you see?

Source: *Instagram*

What objects do you see?



Source: [Instagram]

What objects do you see? Now you can only choose from one of the **COCO* dataset labels**

object 1

object 2

object 3

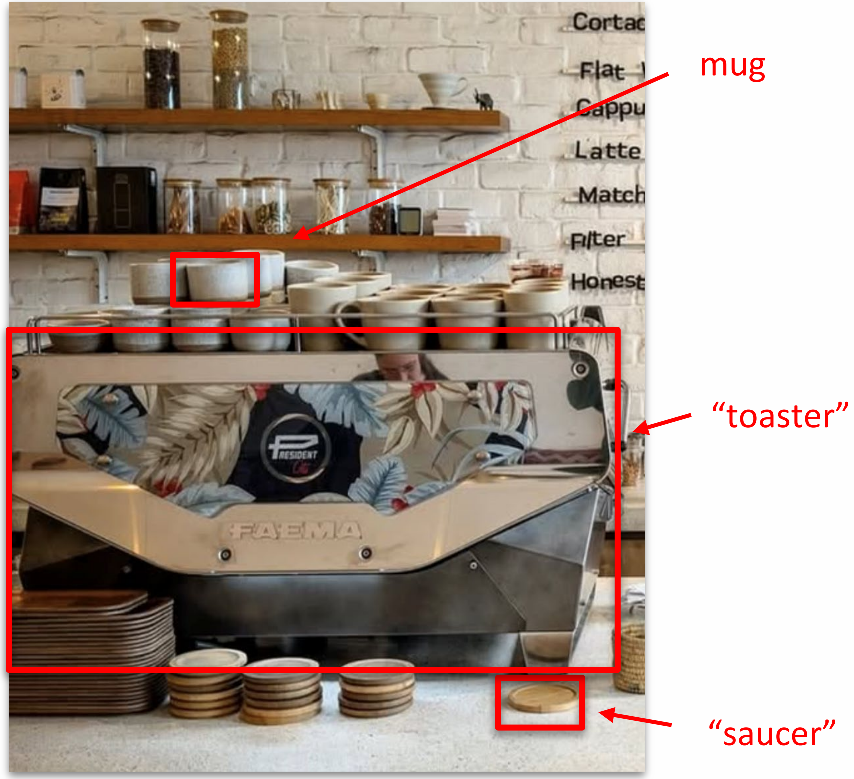
Sample COCO labels

person	wine glass	toaster
bicycle	cup	sink
car	fork	refrigerator
motorcycle	knife	book
airplane	spoon	clock
bus	bowl	vase
train	banana	scissors
truck	apple	teddy bear
boat	sandwich	hair drier
traffic light	orange	toothbrush

**COCO - Common Objects in Context [link]*

Source: Instagram

Problem Statement: Object Detection in the Real World



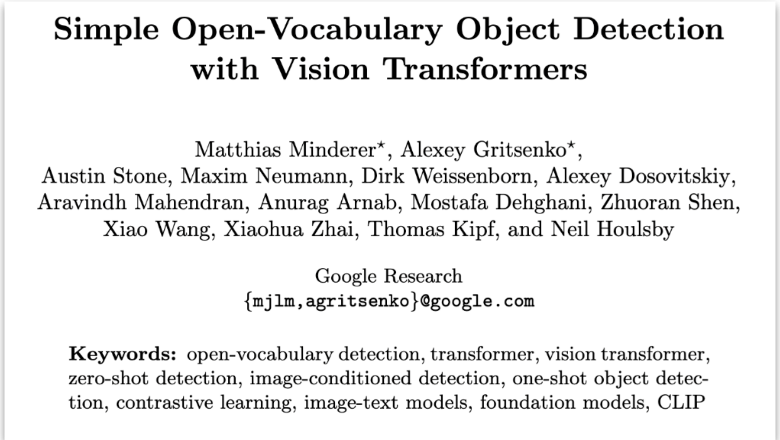
Closed-Vocabulary Detection Problem:

- Models (e.g., COCO, LVIS) are trained on a ***fixed set*** of categories (80, 1,200, etc.)
- ***Out-of-vocabulary*** objects are either ignored or misclassified
- Scaling to cover "every object in the world" with manual labels is ***impossible***

Need: *An object detector that:*

- *Works with **natural language labels** (no fixed class list)*
- *Generalizes to **unseen categories** without retraining*
- *Retains **competitive** performance on **known** categories*

Proposed Solution: OWL-ViT: Vision Transformer for Open-World Localization



Simple Open-Vocabulary Object Detection with Vision Transformers

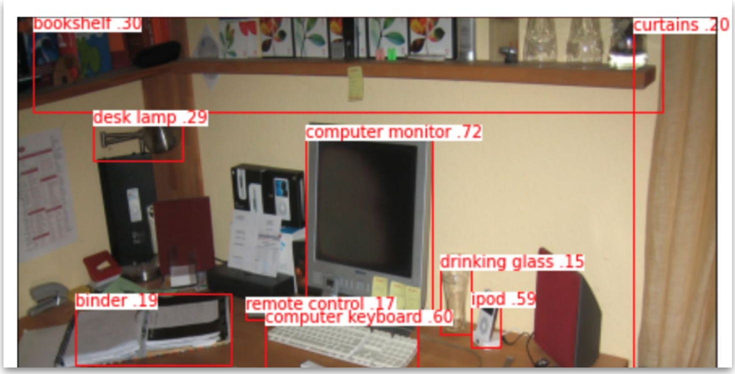
Matthias Minderer*, Alexey Gritsenko*,
Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy,
Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen,
Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby

Google Research
{mjlm,agritsenko}@google.com

Keywords: open-vocabulary detection, transformer, vision transformer, zero-shot detection, image-conditioned detection, one-shot object detection, contrastive learning, image-text models, foundation models, CLIP

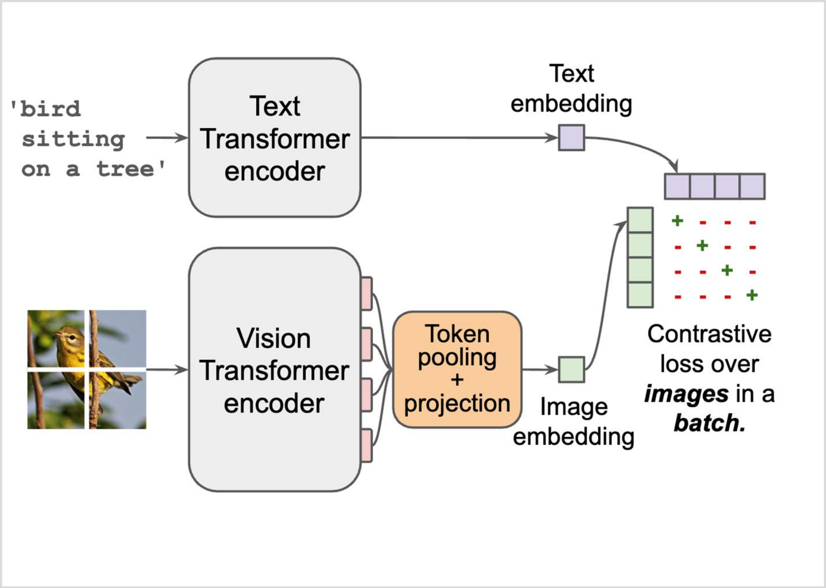
Contributions

- **Open-Vocabulary Detection**: detects objects described in text, not limited to training labels
- **Zero-Shot Generalization**: finds novel categories without retraining (e.g., "espresso machine")
- **Simplicity + Scaling**: large-scale pre-training + ViT + end-to-end fine-tuning outperforms more complex architectures



Approach: Two Stages: **Large-Scale Pre-Training** + Detection Fine-Tuning

Image-level contrastive pre-training

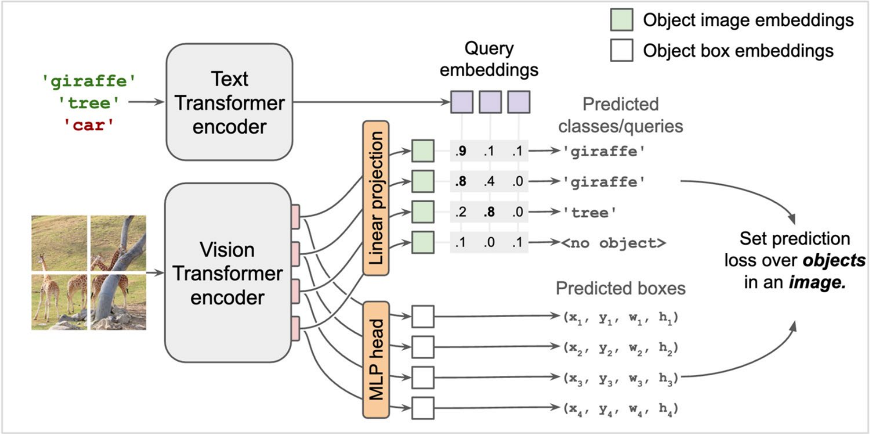


***Stage 1**: Contrastively pre-train image and text encoders on large-scale image-text data*

- ***Vision***:
  - Model: ViT: [B]ase, [L]arge, [H]uge / 16-32 (patch size); R50+H – ResNet50 + ViT-H[uge]
- ***Text***: Transformer with 12 layers & 8 heads
- ***Data***: 3.6 billion image-text pairs; batch size 256
- Both Text and Image encoders are trained *from scratch*

Approach: Two Stages: Large-Scale Pre-Training + **Detection Fine-Tuning**

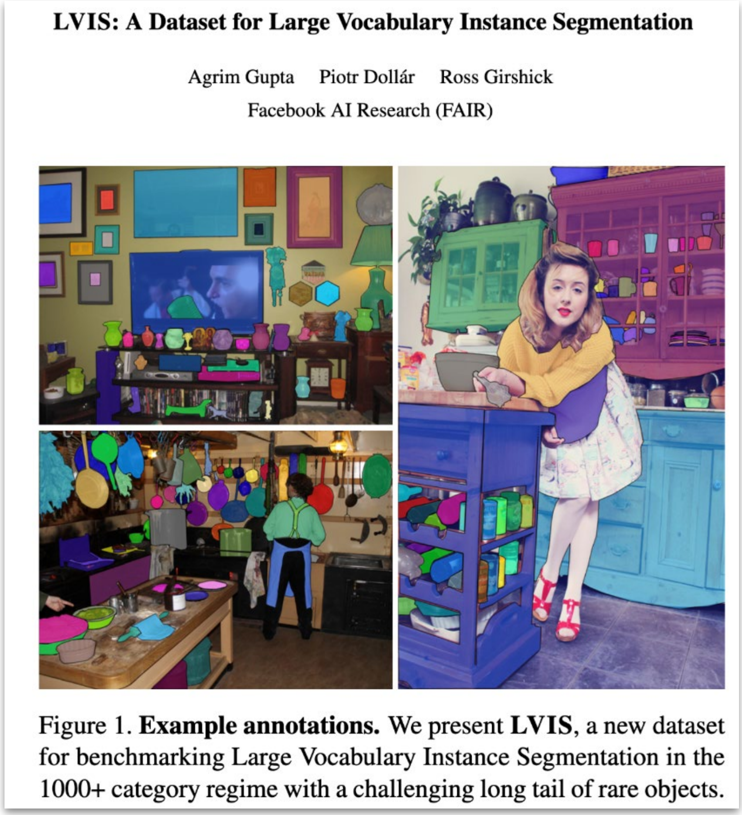
Transfer to open-vocabulary detection



***Stage 2**: Add Detection Heads and fine-tune on medium-sized detection data*

- ***Text:*** Text encoder from CLIP is retained; At inference, user supplies arbitrary text labels (“espresso machine”) → “query embedding”
- ***Vision***:
  - Remove the token pooling + projection layer
  - Linearly project each output token representation to obtain **per-object image embeddings for classification**
  - Max number of predicted objects = number of tokens (576+)
  - Box coordinates come from a separate MLP head
- ***Data***: Medium-scale detection datasets (e.g., LVis, COCO, Objects365)
- **Text** encoder is **frozen**; we’re only retraining the ViT

Data: LVIS – Test-bed for RARE (“unseen”) categories


LVIS: A Dataset for Large Vocabulary Instance Segmentation

Agrim Gupta Piotr Dollár Ross Girshick
Facebook AI Research (FAIR)

Figure 1. **Example annotations.** We present **LVIS**, a new dataset for benchmarking Large Vocabulary Instance Segmentation in the 1000+ category regime with a challenging long tail of rare objects.

LVIS: A Dataset for Large Vocabulary Instance Segmentation [arxiv]

RESULTS: Open-Vocab Detection Performance

Highly competitive results for zero-shot performance (on “unseen” classes)

	Method	Backbone	Image-level	Object-level	Res.	AP^{LVIS}	AP^{LVIS}_{rare}
LVIS base training:							
1	ViLD-ens [12]	ResNet50	CLIP	LVIS base	1024	25.5	16.6
2	ViLD-ens [12]	EffNet-b7	ALIGN	LVIS base	1024	29.3	26.3
3	Reg. CLIP [45]	R50-C4	CC3M	LVIS base	?	28.2	17.1
4	Reg. CLIP [45]	R50x4-C4	CC3M	LVIS base	?	32.3	22.0
5	OWL-ViT (ours)	ViT-H/14	LiT	LVIS base	840	35.3	23.3
6	OWL-ViT (ours)	ViT-L/14	CLIP	LVIS base	840	34.7	25.6

Training: LVIS base (common categories)

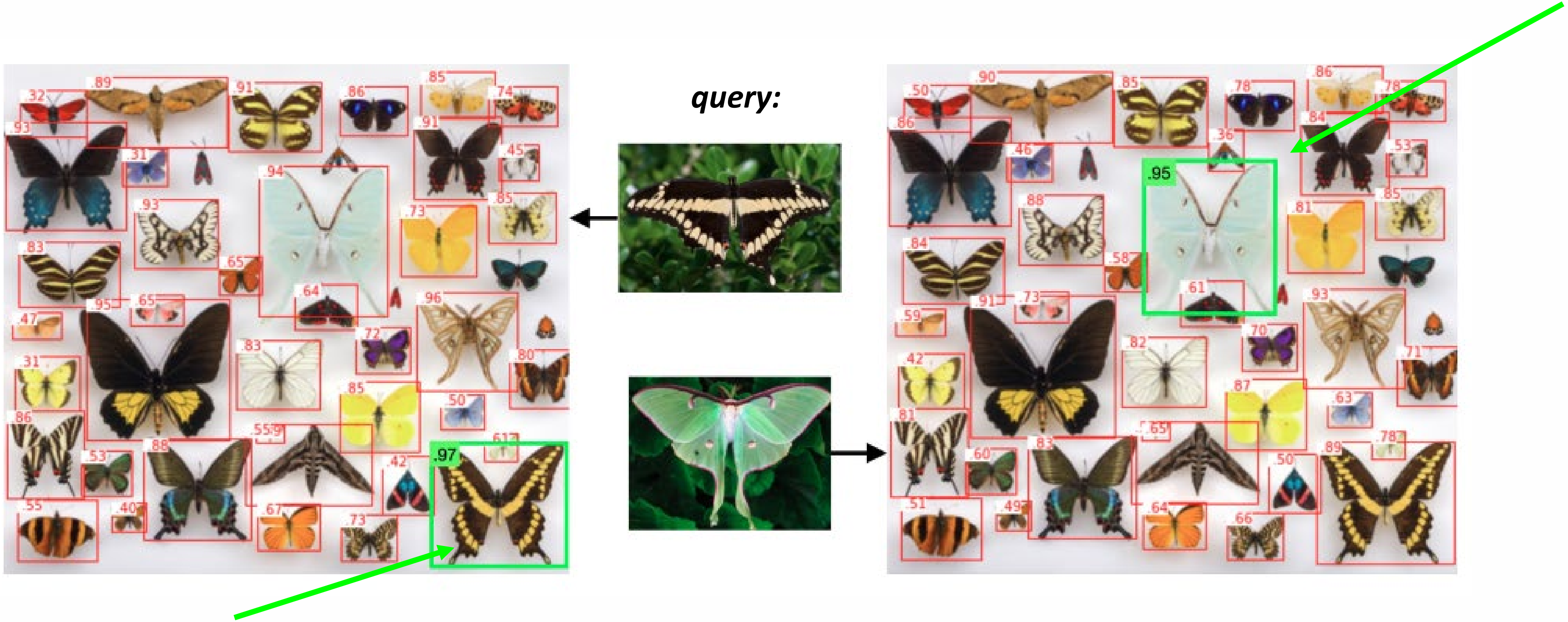
Testing:

- AP^{LVIS} – Precision on ALL categories
- AP_{rare} – Rare (-> *unseen categories*) – basically, zero-shot inference

See this Appendix slide for more info on specific datasets

RESULTS: Image-Conditioned Detection Performance

OWL-ViT strongly outperforms the best task-specific models by a 72% margin



Idea: Use image embeddings (instead of text) to *"query"* the input image and find most relevant objects

Discussion: Loss Functions for Open-Vocabulary Detection

Challenge

- Long-tailed datasets (e.g., LVIS) are federated, not every object is annotated exhaustively
- Objects can have multiple valid labels (e.g., “cup” and “mug”)
- Softmax cross-entropy (pick one label) will penalize reasonable predictions

OWL-ViT Adaptation

- Replace softmax with sigmoid focal loss
- Each class scored independently → allows multiple labels per object
- Focal term helps with imbalance between frequent vs. rare classes

Discussion

- Does this change make evaluation fairer or just easier for the model?
- How do we decide what counts as a “correct” label in open vocab? (cup vs. mug)
- Should we trust model predictions that go beyond what the dataset annotates?

OWLv2: Improving OWL performance by scaling Self-Training

Scaling Open-Vocabulary Object Detection

Matthias Minderer Alexey Gritsenko Neil Houlsby
Google DeepMind
{mjlm, agritsenko, neilhoulsby}@google.com

Abstract

Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (≈10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With a ViT-L/14 architecture, OWL-ST improves AP on LVIS rare classes, *for which the model has seen no human box annotations*, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling. Code and checkpoints are available on GitHub.[1]

Scaling Open-Vocabulary Object Detection [NeurIPS '23]

v1 Limitation: Detection phase has very little data compared to the pre-training phase

v2 Solution:

- OWLv2 uses OWL-ViT to automatically generate **pseudo-labels** (bounding boxes + class labels) on vast web-scraped image–text data; use for noisy supervision
- Go from a few hundred thousand detection examples to **billions**

Results:

- Substantial Gains in Rare-Category Detection:
  - AP_{rare} jumps from 31.2% to ~44.6%

Summary

✅ Strengths

- Open-Vocabulary Detection
  (Text & Image Queries)
- Simple, Modular, and Efficient Architecture
- Scales with data and model size

❌ Weaknesses

- Limited amount of detection data
  (solved in v2)
- Purely discriminative (no captioning)
  (solved in DetCLIP-v3)
- Frozen text encoder limits richness
  (solved in DetCLIP-v3)
- Box precision is only moderate
  (solved in Grounding DINO)

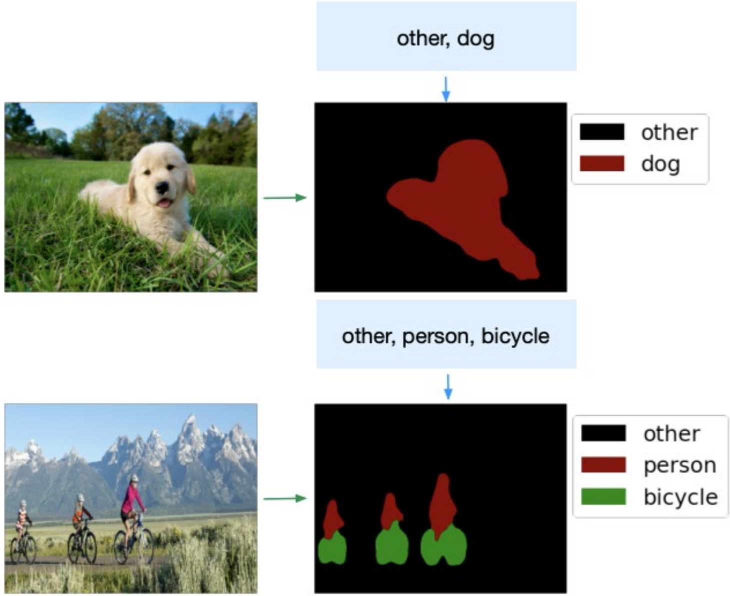
OWL-ViT's role: the proof of concept that contrastive pretrained ViTs can be adapted into open-vocab detectors with almost no architectural changes.
Where it falls short: it's not generative (can't invent labels)

LSeg

Language-driven Semantic Segmentation

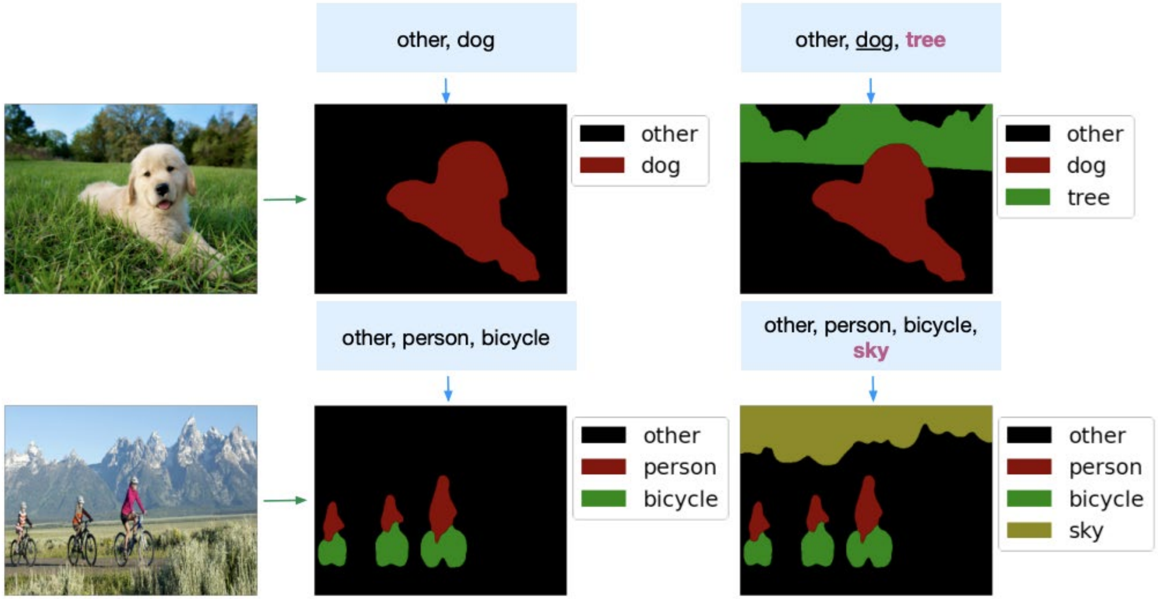
Problem Statement

- CLIP at pixel-level segmentation
- Allows model to potentially learn more precise object recognition



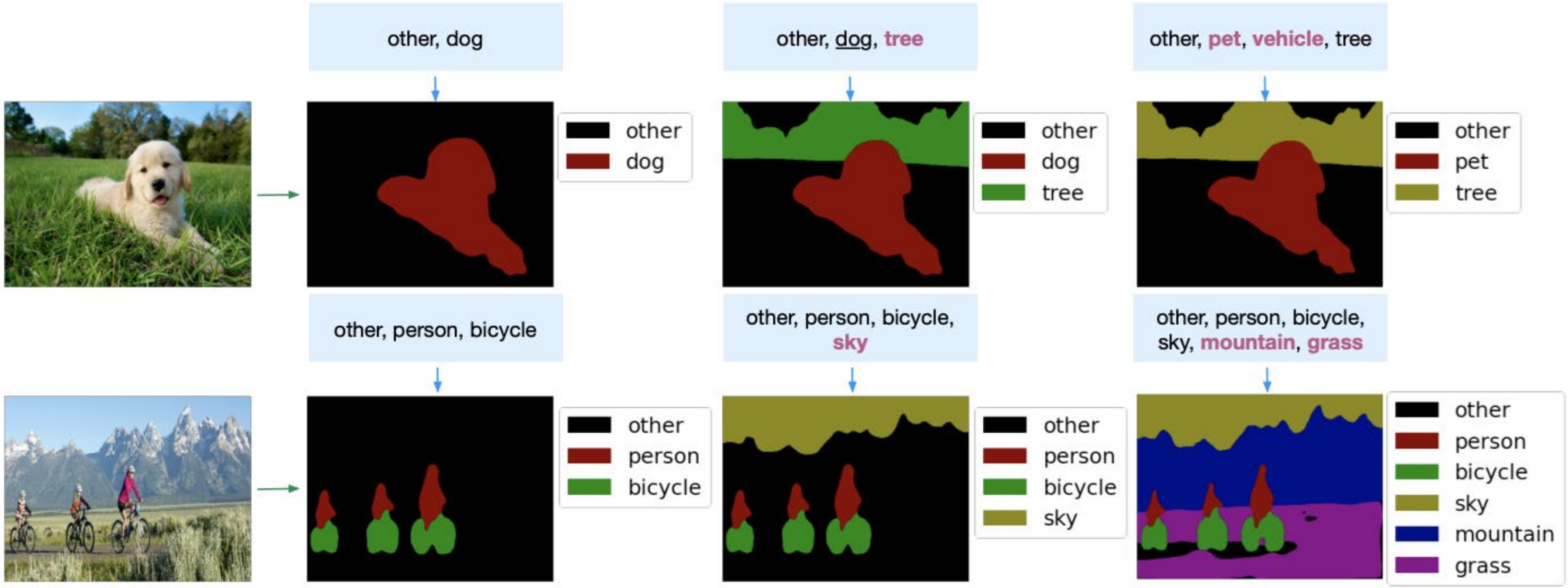
Problem Statement

- CLIP at pixel-level segmentation
- Allows model to potentially learn more precise object recognition

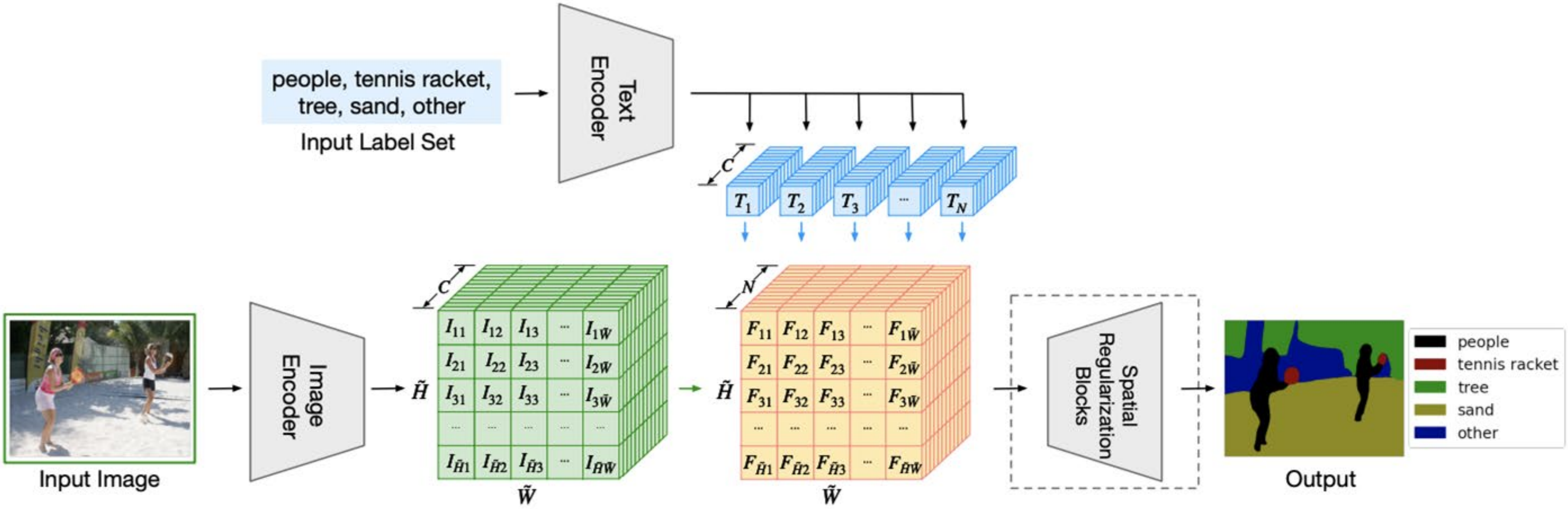


Problem Statement

- CLIP at pixel-level segmentation
- Allows model to potentially learn more precise object recognition



Approach: Architecture

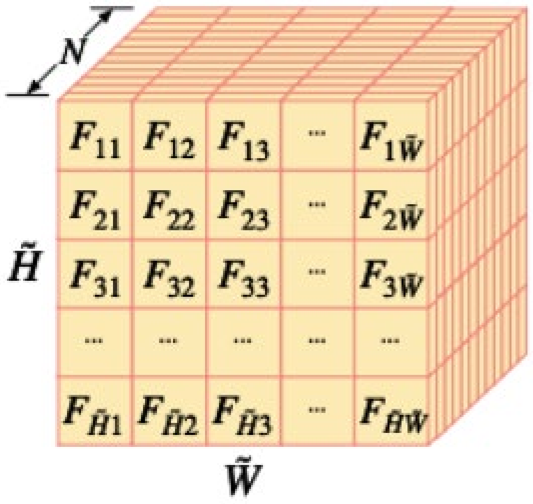


- Text embeddings per input word
- Image embedding per input pixel (after downsampling)

Approach: Contrastive Learning

- Inner product between text and image embeddings
  - Then Softmax (Over what dimension?)

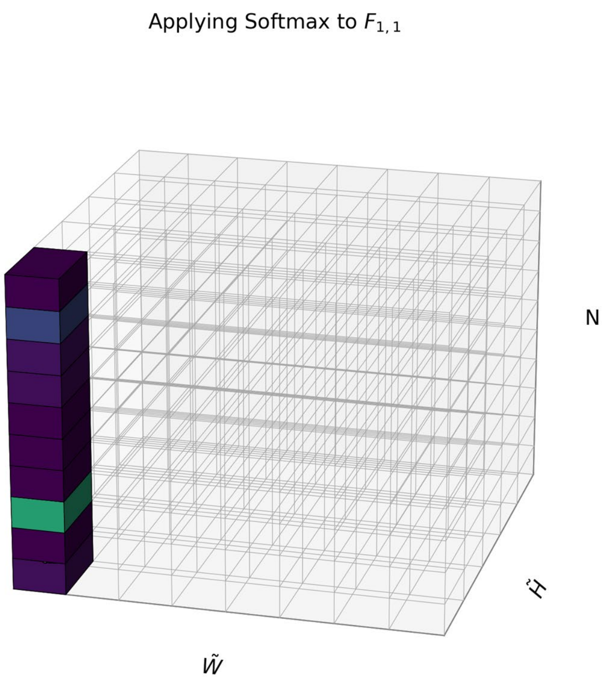
$$f_{ijk} = I_{ij} \cdot T_k.$$



Approach, Contrastive Learning

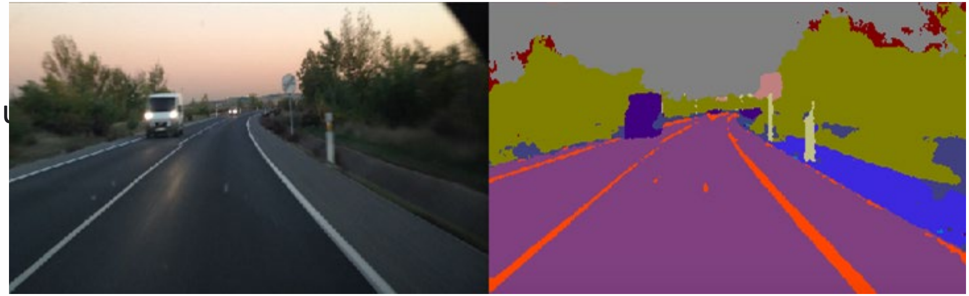
- Softmax over pixels with low temperature (t)
  - Why low temperature?

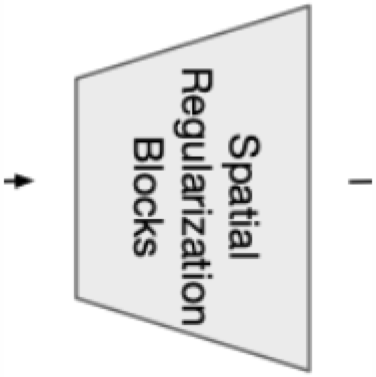
$$\sum_{i,j=1}^{H,W} \text{softmax}_{y_{ij}} \left(\frac{F_{ij}}{t} \right),$$

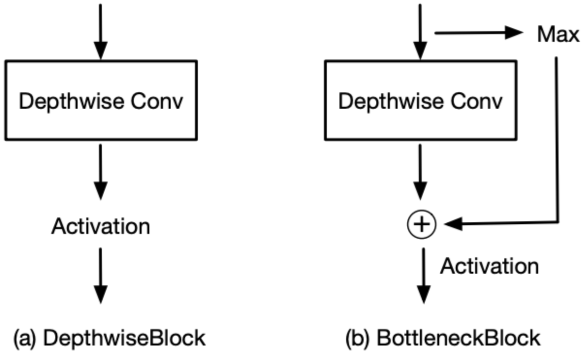


Approach, Spatial Regularization

- Depthwise convolution for regularization
  - Why do regularization at all?
- Then bilinear interpolation to recover original resolu







Experiments and Results

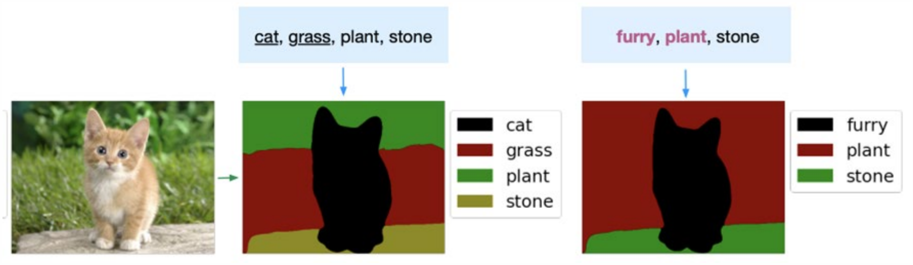
Zero-shot performance matches SOTA one-shot

Model	Backbone	Method	mIoU
OSLSM	VGG16	1-shot	70.3
GNet		1-shot	71.9
FSS		1-shot	73.5
DoG-LSTM		1-shot	80.8
DAN	ResNet101	1-shot	85.2
HSNet		1-shot	86.5
LSeg	ResNet101	zero-shot	84.7
LSeg	ViT-L/16	zero-shot	**87.8**

Table 3: Comparison of mIoU on FSS-1000.

✅ Strengths

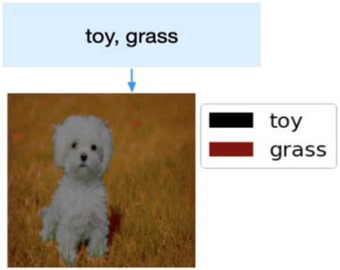
- Embedding training allows for hierarchical knowledge at test time
- Per pixel contrastive loss = tighter prediction boundaries



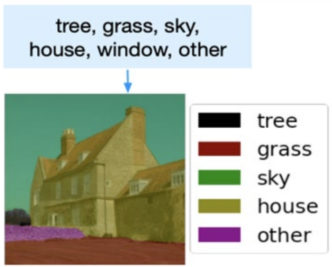
❌ Weaknesses

- Higher memory usage compared to bounding box approach
- Granularity Gap:

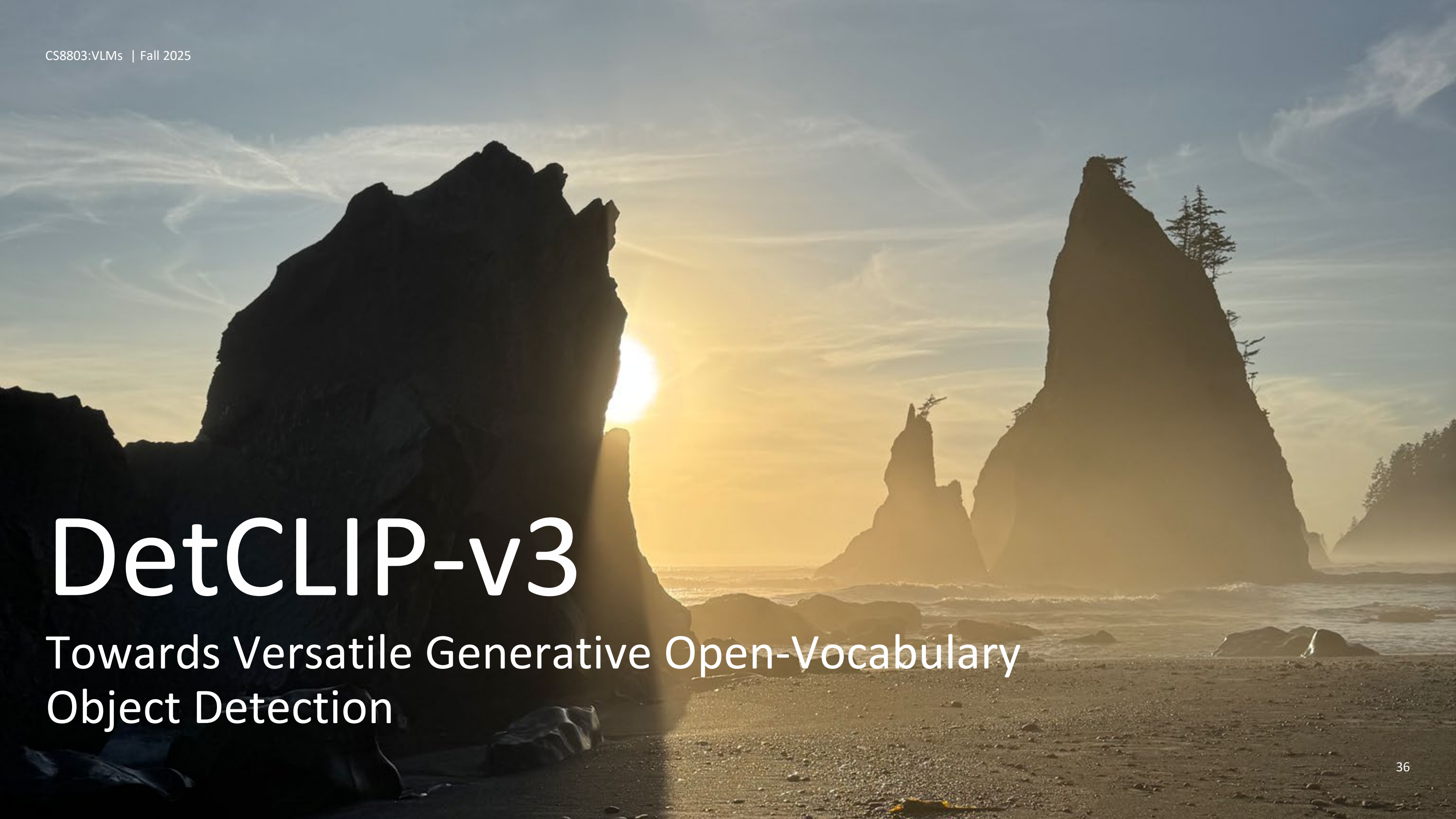
Negative samples missing from training



Only provides one prediction per pixel (could be multiple valid ones)



LSeg demonstrates that you can do **semantic segmentation** without being tied to a fixed class list by **aligning** per-pixel **image embeddings** directly with **language embeddings.**

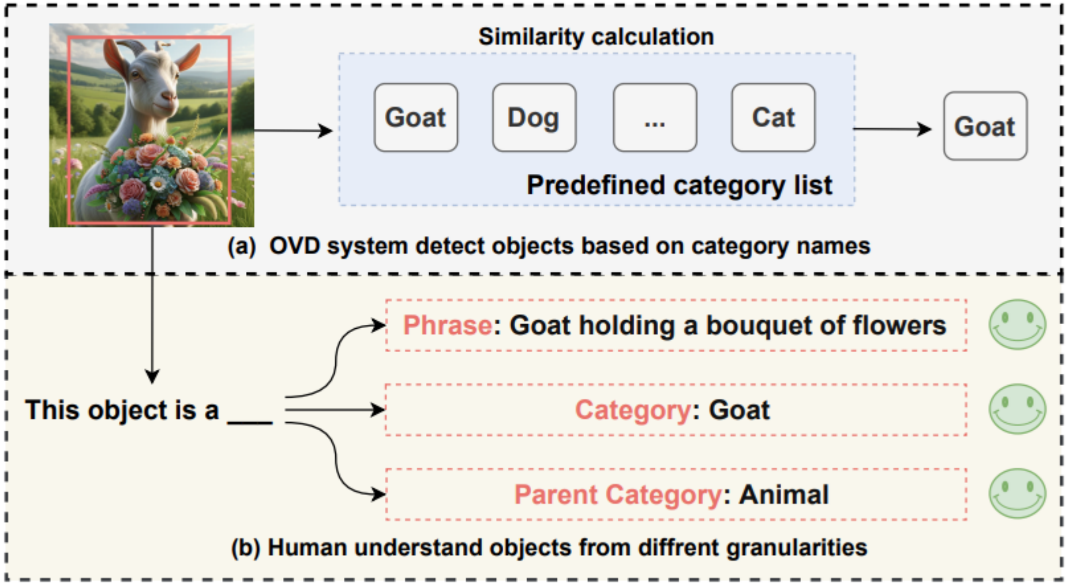


DetCLIP-v3

Towards Versatile Generative Open-Vocabulary Object Detection

DetCLIP-v3: Background & Motivation

- Existing OVD models are limited by their reliance on a **predefined object category list**, which hinders their usage in practical scenarios.
- In contrast, human cognition demonstrates much more versatility. For example, humans are able to understand objects from different **granularities**, in a **hierarchical** manner.

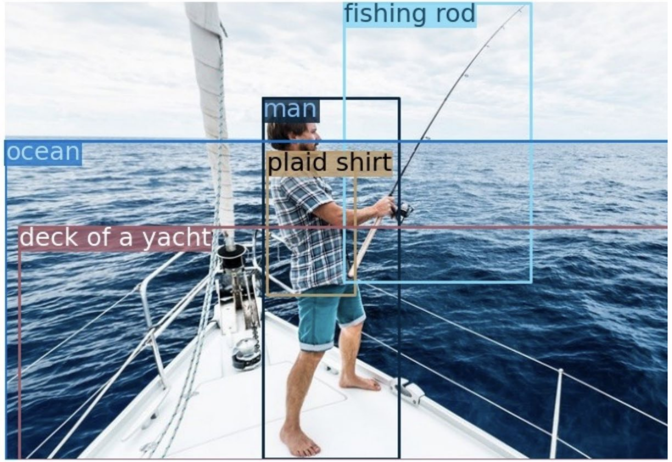


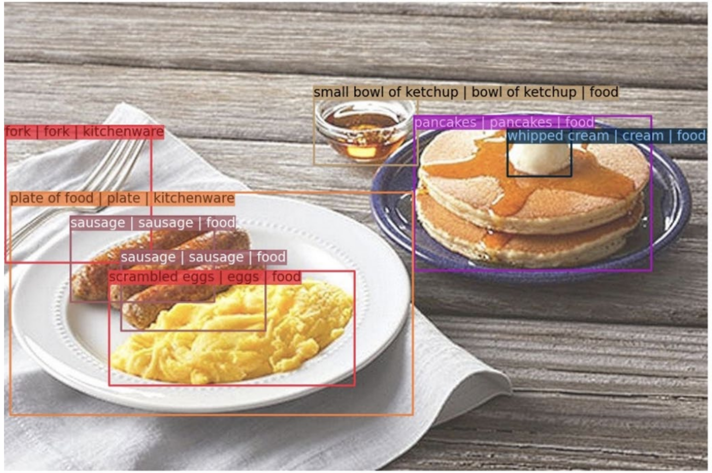
Slide inspired by Lewei Yao

DetCLIP-v3: Overview

Overview

DetCLIPv3 is a high-performing detector that excels not only at open-vocabulary object detection, but also generating hierarchical descriptions for detected objects.



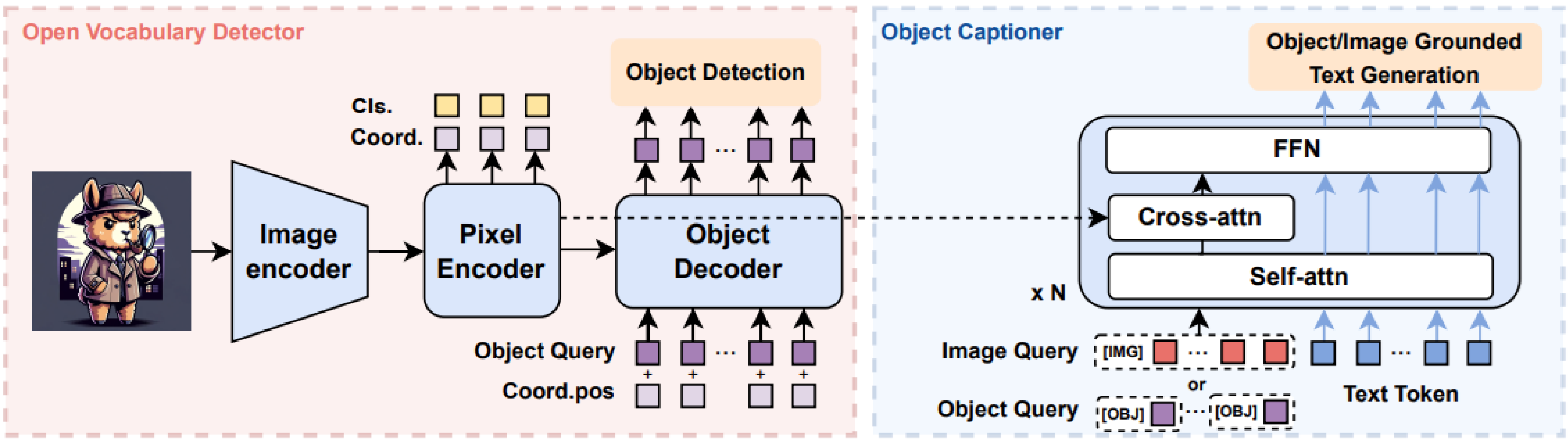


Slide inspired by Lewei Yao

Architecture

Model Architecture

The model is powered by an open-vocabulary object detector, coupled with an object captioner for generating hierarchical and descriptive object concepts.

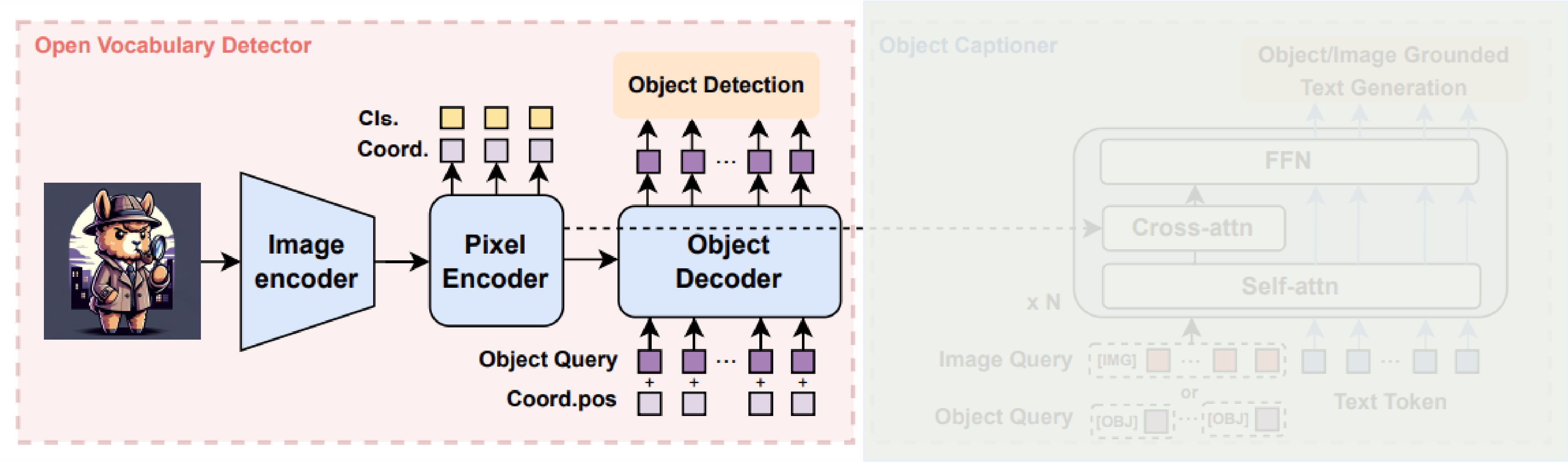


Slide inspired by Lewei Yao

Architecture

Model Architecture

A dual-path model comprising a **visual detector** and **text encoder**
Visual object detector employs a **DETR-like** architecture
Utilizes text features to select the top-k visual tokens from a pixel encoder based on **similarity**

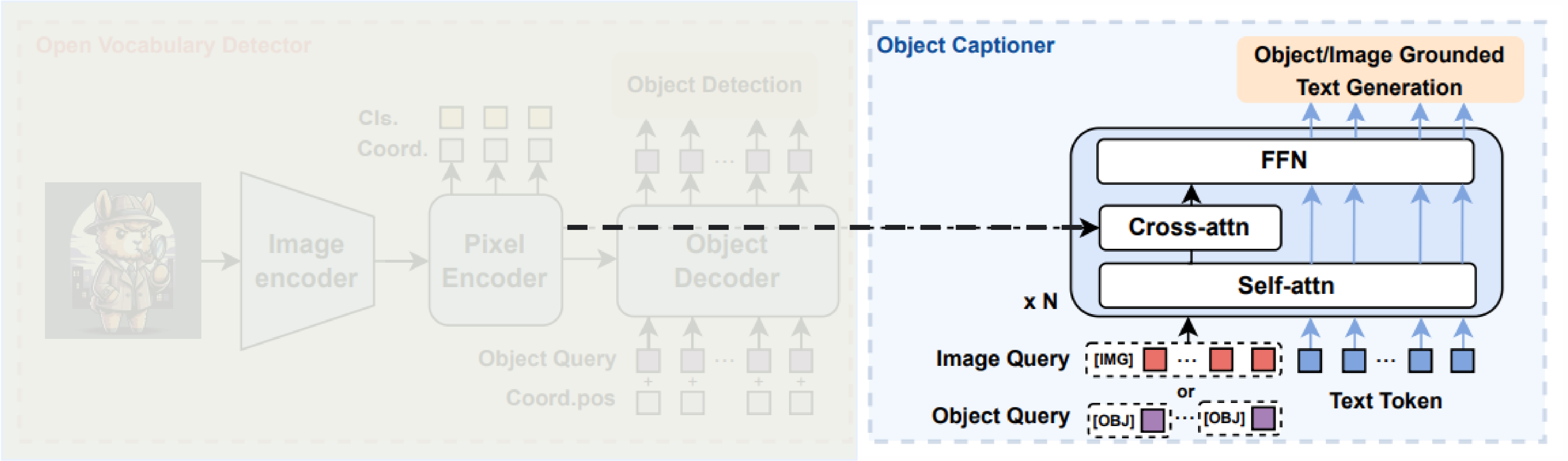


Slide inspired by Lewei Yao

Architecture

Model Architecture

A Transformer-based architecture initialized with the weights of QFormer[1]
2 types of visual queries: **image and object-level** (provided by the OV detector)
Visual queries interact with features from the pixel encoder via **deformable** cross-attention



[1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

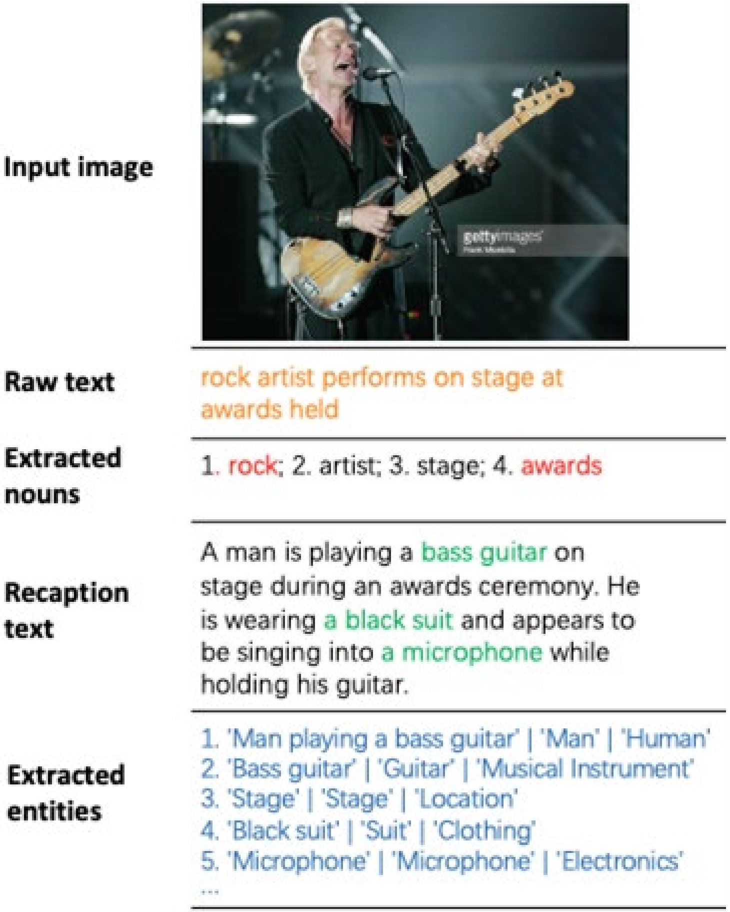
Slide inspired by Lewei Yao

Data

Dataset Construction

To construct a dataset with diverse object-level multi-granular descriptions, an auto-annotation pipeline is developed with 4 steps:

1. Re-captioning image-text pairs with a VLM (InstructBLIP)
2. Entity extraction using GPT-4
3. Fine-tuning the VLM (LLaVA) for large-scale annotation
4. Auto-labeling for bounding boxes



Slide inspired by Lewei Yao

Training

Training Strategy

Learning to generate diverse object-level descriptions requires significant computational resources.

To improve training efficiency, DetCLIPv3 is trained under a 'pretraining + finetuning' paradigm consisting of 3 training stages:

1. **Training the OV detector** with human-annotated datasets (Objects365 + GoldG)
2. **Pretraining the object captioner** (and freeze other parts) using image-text pairs with low resolution input
3. **Holistic finetuning** with all datasets on high resolution inputs. In this stage, all parts of the network are unfrozen and a filtered subset of high-quality, auto-annotated image-text pairs are leveraged for training object-level description generation.

Slide inspired by Lewei Yao

Experiments

DetCLIPv3 achieves SoTA zero-shot OVD performance on a 1203-class dataset LVIS, surpassing previous methods by a large margin.

	Method	Backbone	Pre-training data	LVIS^{minival}			
				AP_{all}	AP_r	AP_c	AP_f
1	GLIP [29]	Swin-T	O365,GoldG,Cap4M	26.0	20.8	21.4	31.0
2	GLIPv2 [65]	Swin-T	O365,GoldG,Cap4M	29.0	–	–	–
3	CapDet [38]	Swin-T	O365,VG	33.8	29.6	32.8	35.5
4	GroundingDINO [36]	Swin-T	O365,GoldG,Cap4M	27.4	18.1	23.3	32.7
5	OWL-ST [43]	CLIP B/16	WebLI2B	34.4	38.3	–	–
6	DetCLIP [58]	Swin-T	O365,GoldG,YFCC1M	35.9	33.2	35.7	36.4
7	DetCLIPv2 [60]	Swin-T	O365,GoldG,CC15M	40.4	36.0	41.7	40.4
8	DetCLIPv3	Swin-T	O365,V3Det,GoldG,GranuCap50M	**47.0**	**45.1**	**47.7**	**46.7**

Table 1. Zero-shot fixed AP on LVIS minival.

Slide inspired by Lewei Yao

Experiments

DetCLIPv3 presents robust generalization to domain shifts. For example, it achieves SoTA performance on the COCO-O dataset.

Method	Backbone	COCO AP	COCO-O AP	Effective Robustness
GLIP [29]	Swin-T	46.1	29	+8.0
DetCLIPv3	Swin-T	**47.2**	**38.5**	**+17.3**
DINO [66]	Swin-L	58.5	42.1	+15.8
DyHead [8]	Swin-L	56.2	35.3	+10.0
GLIP [29]	Swin-L	51.4	48	+24.9
GRiT [56]	ViT-H	60.4	42.9	+15.7
DetCLIPv3	Swin-L	48.5	**48.8**	**+27.0**

Table 4. Distribution shift performance on COCO-O.

Slide inspired by Lewei Yao

Visualization (OVD)

Slide inspired by Lewei Yao

Visualization (object captioning)

Slide inspired by Lewei Yao

Summary

Synthesizing it all

Key Takeaways

- Early work like LSeg established the foundation for language-driven semantic segmentation, enabling zero-shot generalization to new categories.
- Building on these principles, OWL-ViT presented a robust and scalable recipe for open-vocabulary object detection.
- DetCLIPv3 marks a significant shift by introducing generative open-vocabulary object detection.
- Collectively, these advancements demonstrate a **clear progression towards increasingly sophisticated and versatile visual understanding.**

Discussion Points

Questions

1. **Closed vs. Open Vocabulary**: Are there any advantages of having a fixed, closed label set (like COCO's 80 categories)?
2. **Boxes vs Pixels**: Is pixel-level segmentation (LSeg) more useful than bounding boxes (OWL-ViT)?
3. **Text Encoder Fine-tuning**: OWL-ViT froze the text encoder. DetCLIPv3 fine-tuned and even added a caption head. Which strategy is safer for generalization, and which risks overfitting?
4. **Evaluation Metrics**: Current metrics (AP50, AP75, mAP) assume fixed vocabularies. How could we fairly measure success in truly open-vocab models?
5. **Applications & Safety**: In our coffee shop example, would you trust OWL-ViT to detect allergens (e.g., "peanut butter jar")? What about rare but critical safety items (e.g., "fire extinguisher")?

Appendix

Datasets [Detailed]

General Object Detection Benchmarks

- COCO (Common Objects in Context)
  ~118k training images, 80 object categories, dense annotations (boxes, masks, captions).
- LVIS (Large Vocabulary Instance Segmentation)
  Extension of COCO with 1,200+ categories, long-tailed distribution. Perfect for open-vocab detection.
- PASCAL VOC
  20 categories, ~10k images. Mostly a “legacy” benchmark.
- Objects365
  ~365 categories, 600k images, large-scale detection dataset
- OpenImagesV4
  Very large-scale (~9M images, 600+ categories), weakly and sparsely annotated bounding boxes.
- V3Det
  Chinese open-domain detection dataset (~13M boxes, ~13k categories)

Dense & Structured Annotations

- **Visual Genome**
  ~100k images with dense region descriptions, attributes, relationships. Often used to link vision with language beyond flat labels (captioning, grounding).
- **FSS-1000** (Few-Shot Segmentation 1000)
  1,000 categories with only a few annotated examples per category. Tailored for few-shot segmentation and open-vocab generalization tests.

Specialized / Custom Datasets

- **GoldG**
  A curated grounding dataset (image–text pairs with region annotations). Smaller but high-quality for grounding tasks.
- **GranuCap50M**
  Large-scale caption dataset with granular, multi-level labels (auto-generated). Used to train DetCLIPv3 for hierarchical captions
- **Custom DetCLIP-v3 Dataset**
  The authors’ auto-annotated mixture: leverages visual LLMs to refine captions, generating rich multi-granular supervision for detection + captioning.

Related Work

Open-Vocab Object Detection

Detects any object described by a text vocabulary

The Gap: Requires a predefined list of categories to search for

DetCLIPv3: Generates rich, hierarchical labels for objects without needing a predefined list

Dense Captioning

Generates text descriptions for specific regions in an image

The Gap: Can only describe a range of visual concepts

DetCLIPv3: Taps into image-text pairs to describe a much wider, diverse range of concepts

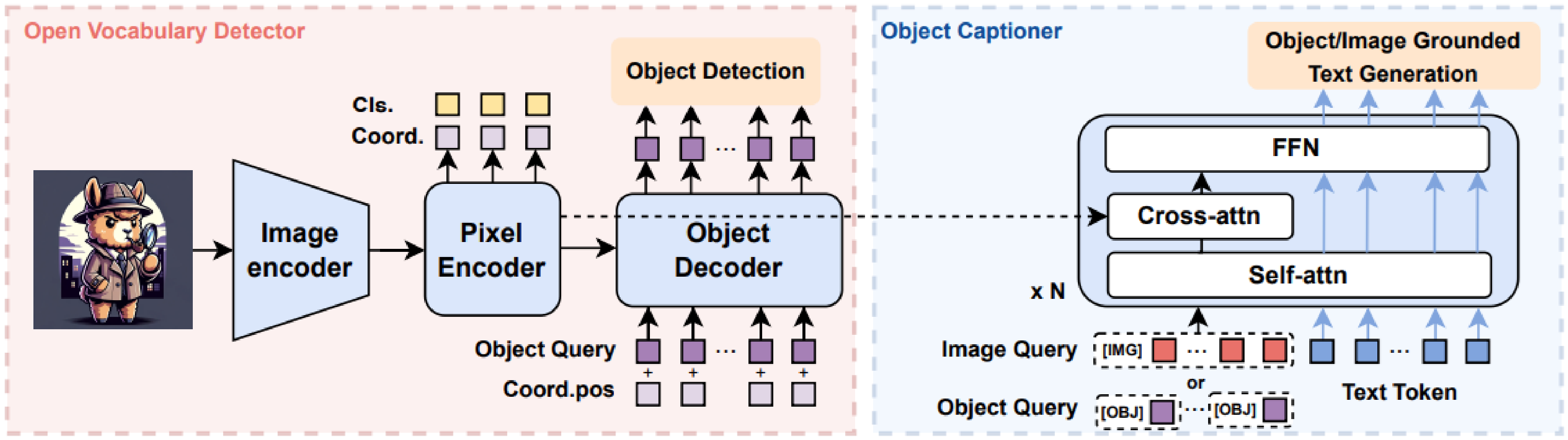
Re-captioning for Better Data

A technique to refine noisy, low-quality image-text data

The Gap: Helps many visual tasks, but OVD potential underexplored

DetCLIPv3: Auto-annotation pipeline to train generative object detector

Approach
Model Design



(left) OV detector localizes objects by category and proposes regions;
(right) captioner assigns hierarchical labels and produces image-level descriptions.

Approach

Dataset Construction

Input image				
Raw text	rock artist performs on stage at awards held	8 Questions To Consider Before Meeting With A Home Designer Fox News	the Woodward's Windows	blonde labrador retriever in snow on a shoot day
Extracted nouns	1. rock; 2. artist; 3. stage; 4. awards	1. questions; 2. meeting; 3. home; 4. designer; 5. fox; 6. news	1. woodward; 2. windows	1. labrador; 2. snow; 3. shoot; 4. day
Recaption text	A man is playing a bass guitar on stage during an awards ceremony. He is wearing a black suit and appears to be singing into a microphone while holding his guitar.	The image depicts a spacious kitchen with wooden cabinets, countertops, and appliances. There is a large island in the center. The kitchen also features a stainless steel refrigerator, oven, and dishwasher ...	The image features a Christmas-themed display in a store window, showcasing a variety of decorations and figurines. There are several mannequins dressed in Victorian-style clothing. Additionally, there are various Christmas trees and wreaths ...	The image features a blonde labrador retriever standing in the snow, looking up and away from the camera. The dog's head is tilted slightly to the side.
Extracted entities	1. 'Man playing a bass guitar' | 'Man' | 'Human' 2. 'Bass guitar' | 'Guitar' | 'Musical Instrument' 3. 'Stage' | 'Stage' | 'Location' 4. 'Black suit' | 'Suit' | 'Clothing' 5. 'Microphone' | 'Microphone' | 'Electronics' ...	1. 'Spacious kitchen' | 'kitchen' | 'Rooms in a house' 2. 'Wooden cabinets' | 'cabinets' | 'Furniture' 3. 'Countertops' | 'countertops' | 'Kitchen appliances' 4. 'Appliances' | 'appliances' | 'Kitchen appliances' 5. 'Large island' | 'island' | 'Kitchen furniture' ...	1. 'Christmas-themed display' | 'display' | 'Store Items' 2. 'Store window' | 'window' | 'Building Parts' 3. 'Figurines' | 'figurines' | 'Decorative Items' 4. 'Several mannequins' | 'mannequins' | 'Store Items' 5. 'Mannequins dressed in Victorian-style clothing' | 'mannequins' | 'Store Items' ...	1. 'Blonde labrador retriever' | 'labrador retriever' | 'Dog breeds' 2. 'Snow' | 'Snow' | 'Weather conditions' 3. 'Dog's head' | 'Head' | 'Body parts'

Illustration of quality issues existing in image-text pair data

Approach

Multi-stage Training Scheme

Dataset Pipeline

1. Re-captioning with VLLM
2. Entity Extraction using GPT-4
3. Instruction tuning of VLLM for large-scale annotation

Pretraining + Finetuning Paradigm

1. Training the OV detector
2. Pretraining the object captioner
3. Holistic finetuning

Strengths

Versatile Generative Open-Vocabulary Detection

State-of-the-Art Performance

Robustness to Distribution Shifts and High Transferability

Efficient and Innovative Architecture & Training

Weaknesses

Incomplete Evaluation Benchmarks for Generative Capabilities

Current Lack of Instruction Control in Detection

Complexity and Cost of Data Auto-Annotation Pipeline

Balancing Performance and Training Efficiency

Lseg: Experiments and Results

- Zero-shot performance matches SOTA one-shot

Model	Backbone	Method	5^0	5^1	5^2	5^3	mean	FB-IoU
OSLSM	VGG16	1-shot	33.6	55.2	40.9	33.5	40.8	61.3
co-FCN	VGG16	1-shot	36.7	50.6	44.9	32.4	41.1	60.1
AMP-2	VGG16	1-shot	41.9	50.2	46.7	34.7	43.4	61.9
PANet	ResNet50	1-shot	44.0	57.5	50.8	44.0	49.1	-
PGNet	ResNet50	1-shot	56.0	66.9	50.6	50.4	56.0	69.9
FWB	ResNet101	1-shot	51.3	64.5	56.7	52.2	56.2	-
PPNet	ResNet101	1-shot	52.7	62.8	57.4	47.7	55.2	70.9
DAN	ResNet101	1-shot	54.7	68.6	57.8	51.6	58.2	71.9
PFENet	ResNet101	1-shot	60.5	69.4	54.4	55.9	60.1	72.9
RePRI	ResNet101	1-shot	59.6	68.6	**62.2**	47.2	59.4	-
HSNet	ResNet101	1-shot	**67.3**	**72.3**	62.0	**63.1**	**66.2**	**77.6**
SPNet	ResNet101	zero-shot	23.8	17.0	14.1	18.3	18.3	44.3
ZS3Net	ResNet101	zero-shot	40.8	39.4	39.3	33.6	38.3	57.7
LSeg	ResNet101	zero-shot	**52.8**	**53.8**	**44.4**	**38.5**	**47.4**	**64.1**
LSeg	ViT-L/16	zero-shot	**61.3**	**63.6**	**43.1**	**41.0**	**52.3**	**67.0**

Table 1: Comparison of mIoU and FB-IoU (higher is better) on PASCAL-5^i.

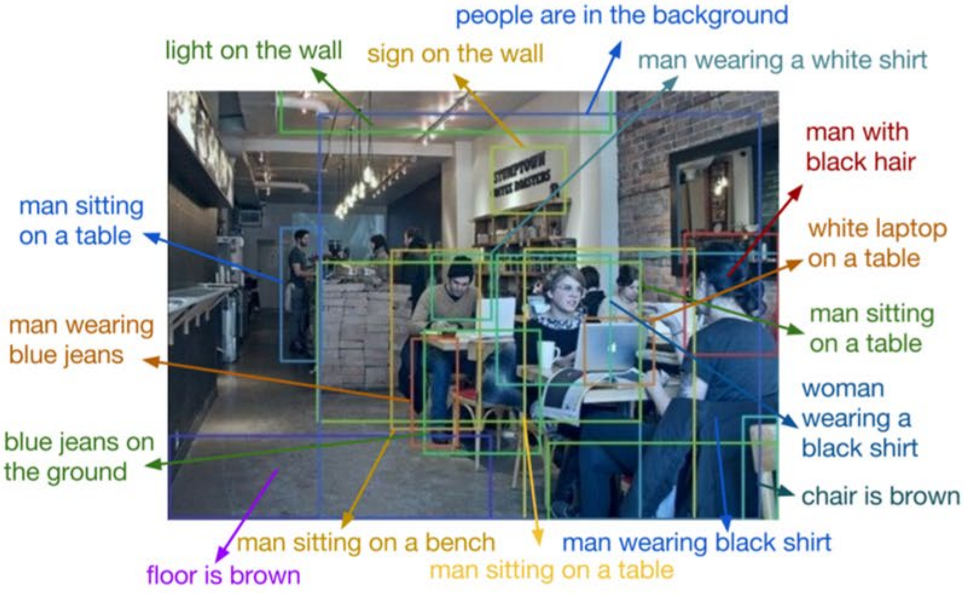
Model	Backbone	Method	mIoU
OSLSM	VGG16	1-shot	70.3
GNet	VGG16	1-shot	71.9
FSS	VGG16	1-shot	73.5
DoG-LSTM	VGG16	1-shot	80.8
DAN	ResNet101	1-shot	85.2
HSNet	ResNet101	1-shot	86.5
LSeg	ResNet101	zero-shot	84.7
LSeg	ViT-L/16	zero-shot	**87.8**

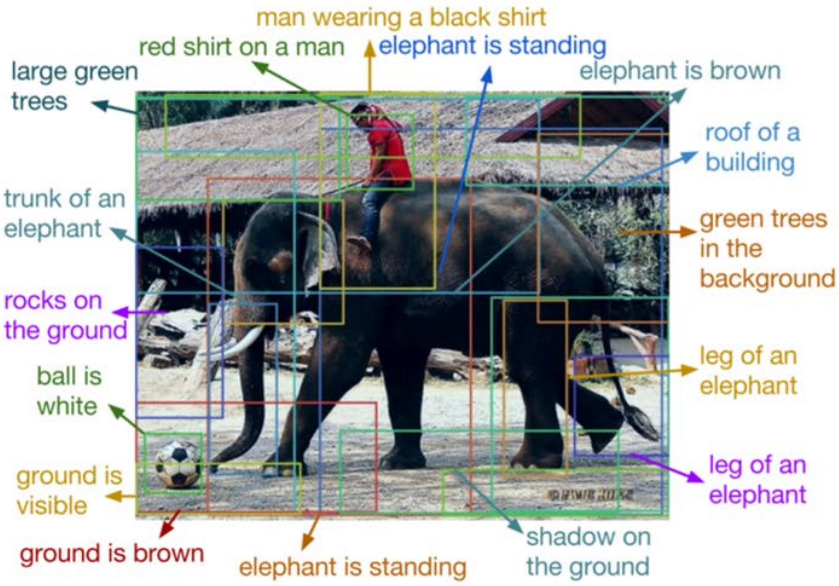
Table 3: Comparison of mIoU on FSS-1000.

Model	Backbone	Method	20^0	20^1	20^2	20^3	mean	FB-IoU
PPNet	ResNet50	1-shot	28.1	30.8	29.5	27.7	29.0	-
PMM	ResNet50	1-shot	29.3	34.8	27.1	27.3	29.6	-
RPMM	ResNet50	1-shot	29.5	36.8	28.9	27.0	30.6	-
RePRI	ResNet50	1-shot	32.0	38.7	32.7	33.1	34.1	-
FWB	ResNet101	1-shot	17.0	18.0	21.0	28.9	21.2	-
DAN	ResNet101	1-shot	-	-	-	-	24.4	62.3
PFENet	ResNet101	1-shot	36.8	41.8	38.7	36.7	38.5	63.0
HSNet	ResNet101	1-shot	**37.2**	**44.1**	**42.4**	**41.3**	**41.2**	**69.1**
ZS3Net	ResNet101	zero-shot	18.8	20.1	24.8	20.5	21.1	55.1
LSeg	ResNet101	zero-shot	**22.1**	**25.1**	**24.9**	**21.5**	**23.4**	**57.9**
LSeg	ViT-L/16	zero-shot	**28.1**	**27.5**	**30.0**	**23.2**	**27.2**	**59.9**

Table 2: Comparison of mIoU and FB-IoU (higher is better) on COCO-20^i.

Open vocabulary tasks





Source: UMich CV Lecture Slides

RESULTS: Open-Vocab Detection Performance

Highly competitive results for zero-shot performance (on "unseen" classes)

	Method	Backbone	Image-level	Object-level	Res.	AP^{LVIS}	AP^{LVIS}_{rare}
	LVIS base training:						
1	ViLD-ens [12]	ResNet50	CLIP	LVIS base	1024	25.5	16.6
2	ViLD-ens [12]	EffNet-b7	ALIGN	LVIS base	1024	29.3	26.3
3	Reg. CLIP [45]	R50-C4	CC3M	LVIS base	?	28.2	17.1
4	Reg. CLIP [45]	R50x4-C4	CC3M	LVIS base	?	32.3	22.0
5	OWL-ViT (ours)	ViT-H/14	LiT	LVIS base	840	35.3	23.3
6	OWL-ViT (ours)	ViT-L/14	CLIP	LVIS base	840	34.7	25.6
	Unrestricted open-vocabulary training:						
7	GLIP [26]	Swin-T	Cap4M	O365, GoldG, ...	?	17.2	10.1
8	GLIP [26]	Swin-L	CC12M, SBU	OI, O365, VG, ...	?	26.9	17.1
9	OWL-ViT (ours)	ViT-B/32	LiT	O365, VG	768	23.3	19.7
11	OWL-ViT (ours)	R26+B/32	LiT	O365, VG	768	25.7	21.6
10	OWL-ViT (ours)	ViT-B/16	LiT	O365, VG	768	26.7	23.6
12	OWL-ViT (ours)	ViT-L/16	LiT	O365, VG	768	30.9	28.8
13	OWL-ViT (ours)	ViT-H/14	LiT	O365, VG	840	33.6	30.6
14	OWL-ViT (ours)	ViT-B/32	CLIP	O365, VG	768	22.1	18.9
15	OWL-ViT (ours)	ViT-B/16	CLIP	O365, VG	768	27.2	20.6
16	OWL-ViT (ours)	ViT-L/14	CLIP	O365, VG	840	34.6	31.2

Table 1

Training: LVIS base (common categories)

Testing:

- AP^{LVIS} – Precision on ALL categories
- AP_{rare} – Rare (-> *unseen categories*) – basically, zero-shot inference

Training: O365 (Objects365) + VG (Visual Genome)

See this Appendix slide for more info on specific datasets