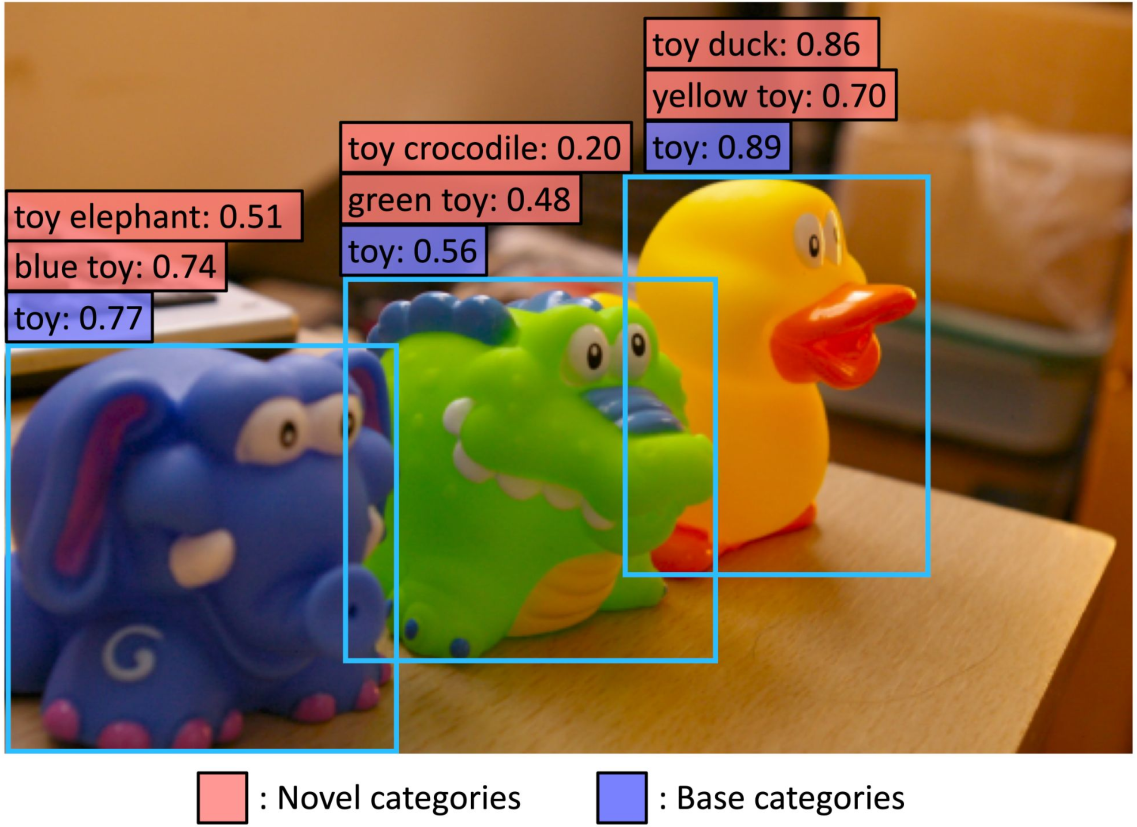
Zero-Shot Detection via Vision and Language Knowledge Distillation

ViLD: **Vi**sion and **L**anguage Knowledge **D**istillation

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, **Yin Cui**
Google Research

MSR V+L Summer Talk Series, July 28, 2021

Zero-shot open vocabulary detection



Motivation

Dataset collection for large vocabulary detection

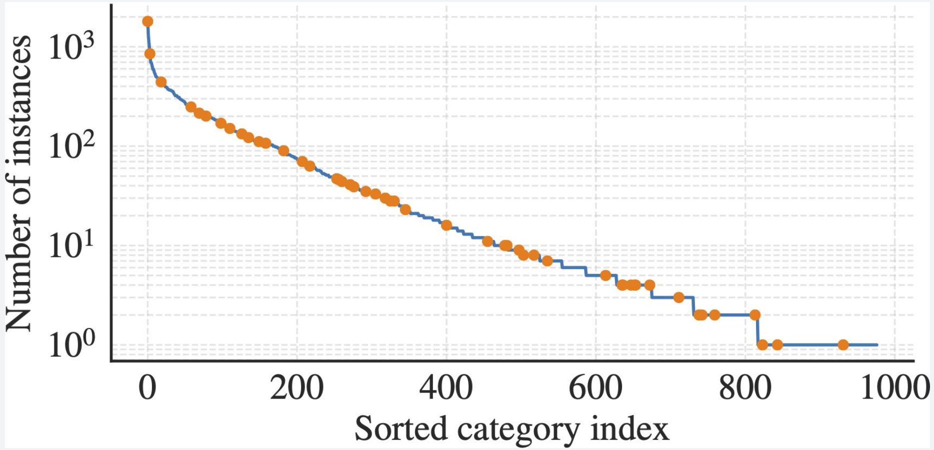
Dataset	# images	# boxes	# categories
Pascal VOC	11.5k	27k	20
COCO	159k	896k	80
Objects 365	1800k	29,000k	365
LVIS v1.0	159k	1,514k	**1203**

Google

Long-tailed distribution
Zipf's Law

- Natural object categories follow a long-tailed distribution.
- Exponentially more data is needed for rare categories.
- Expensive to scale up dataset vocabularies.
- Alternatives?

Number of instances per category in LVIS dataset

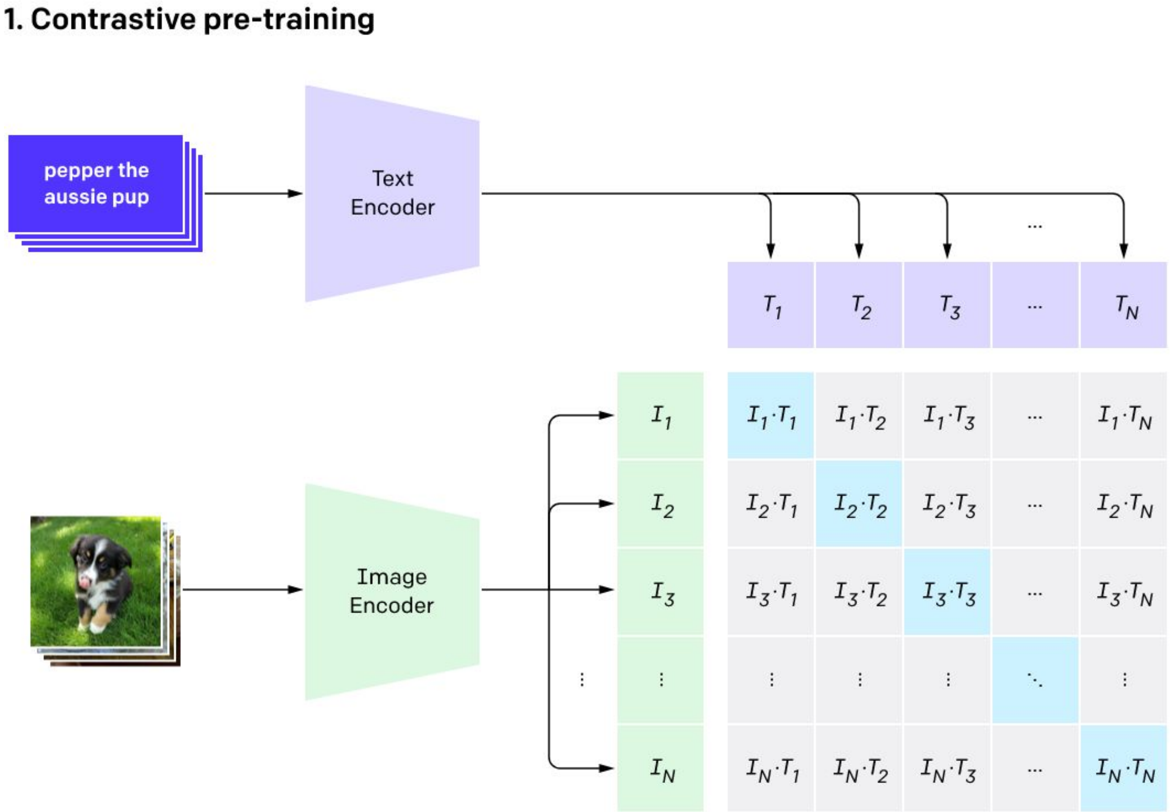


Gupta et al., 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation.

Open vocabulary detection can be a new direction for large vocabulary detection

Recent zero-shot classification models

- CLIP (OpenAI)
- 400M image-text pairs

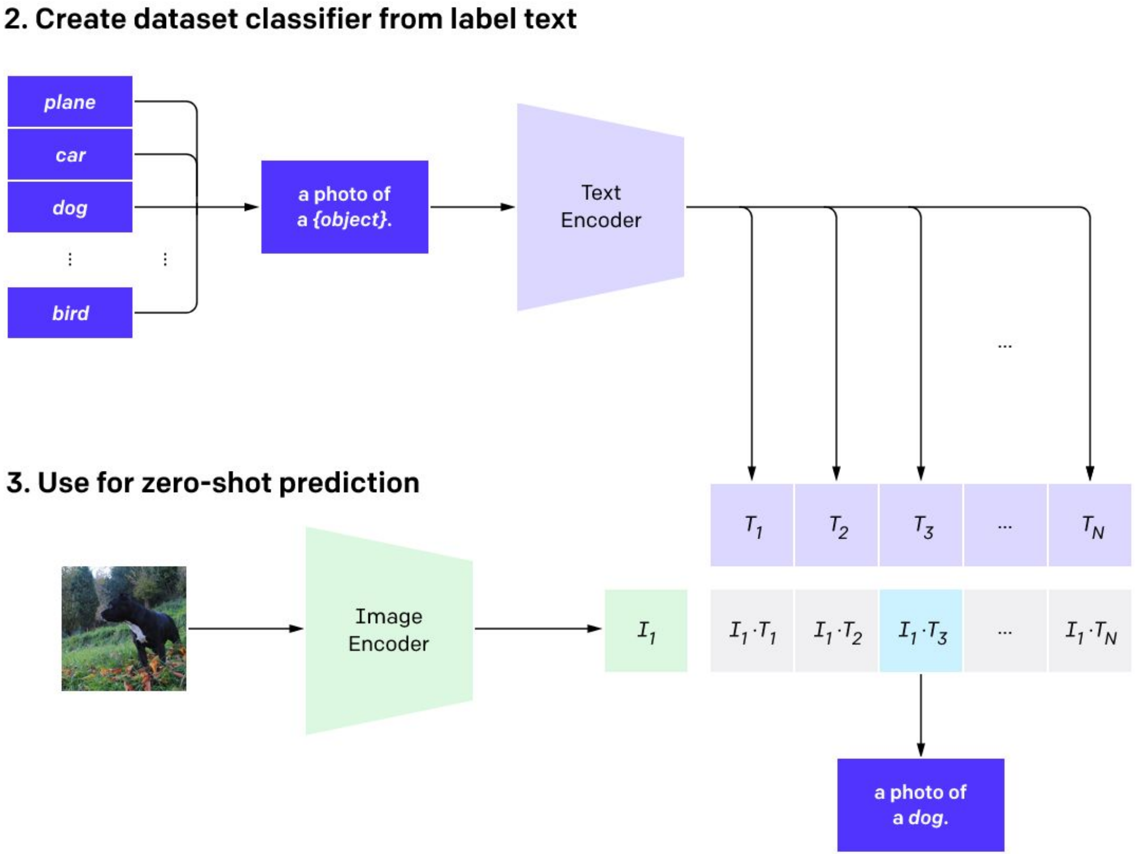


https://openai.com/blog/clip/

Radford et al., 2021. Learning Transferable Visual Models From Natural Language Supervision.

Recent zero-shot classification models

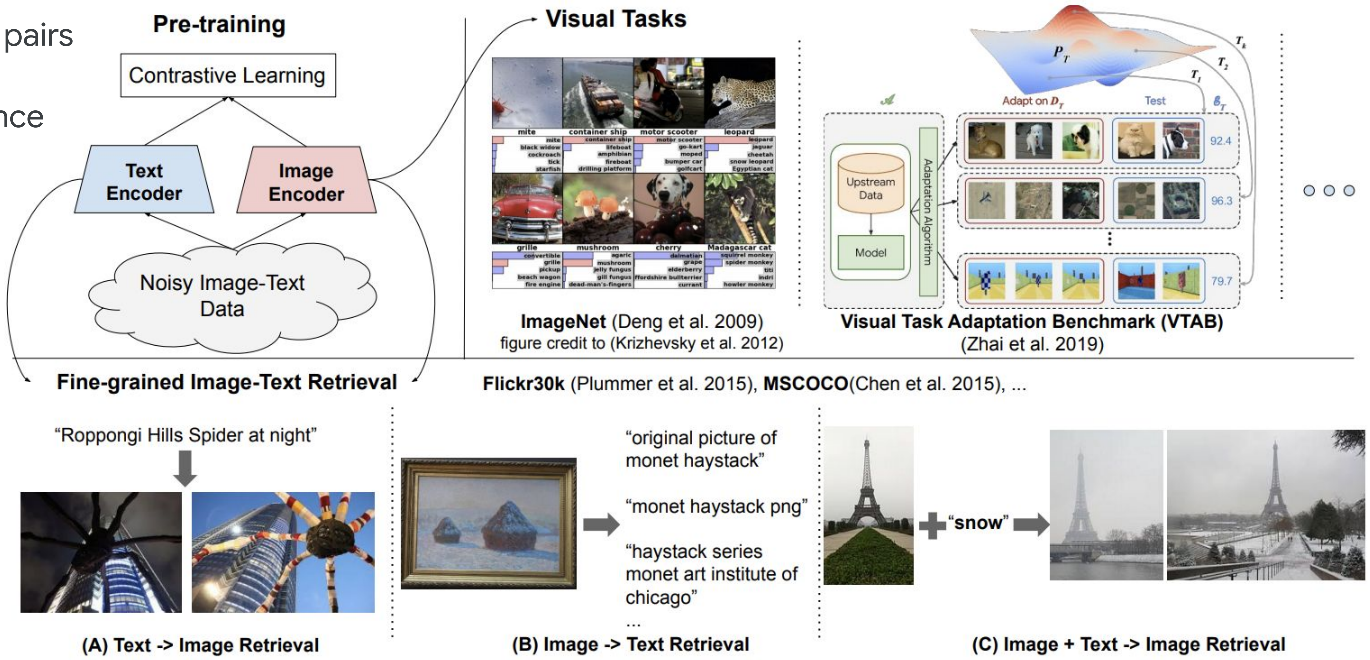
- CLIP (OpenAI)
- 76.2% Top-1 Acc on ImageNet



https://openai.com/blog/clip/
Radford et al., 2021. Learning Transferable Visual Models From Natural Language Supervision.

Recent zero-shot classification models

- ALIGN (Google)
- 1.8B noisier image-text pairs
- Comparable performance with CLIP



Borrowing the knowledge from zero-shot classification model for zero-shot detection

Google

Method

Settings for zero-shot detection

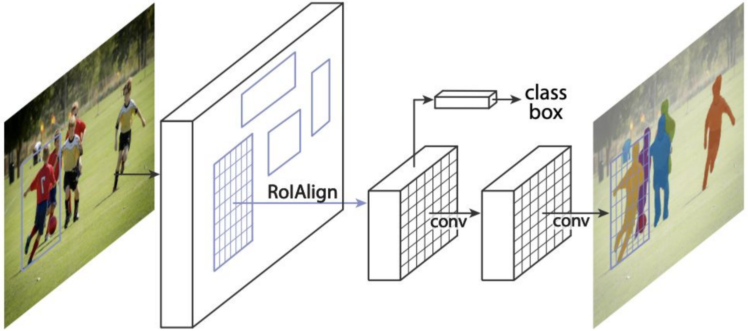
- **Base categories** the model can be trained on.
- **Novel categories** that are never seen during training.
- **Goal**: achieve good performance on novel categories while maintaining performance on base categories.

Note: Our method is zero-shot w.r.t. the detection dataset. Our method does not learn from detection annotations of novel categories.

However, similar concepts of novel categories could be seen in the pre-trained zero-shot classification models (e.g., CLIP)

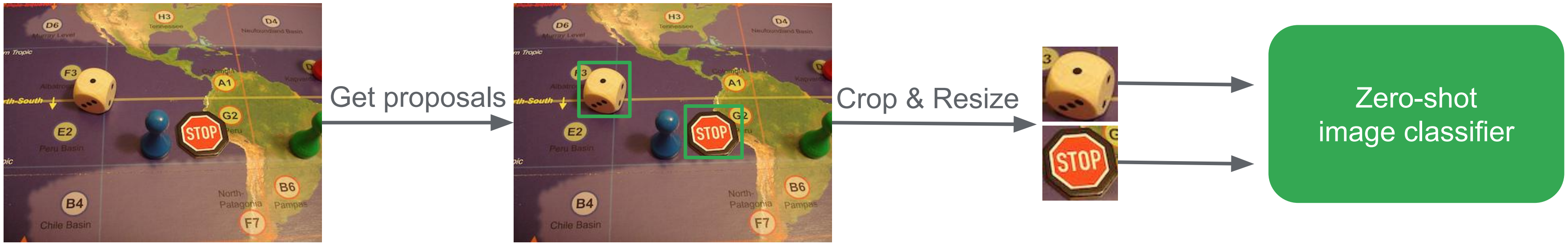
Object proposals for novel categories

- Two-stage object detector (e.g., Mask R-CNN).
- *Class-agnostic* bbox regression and mask prediction.



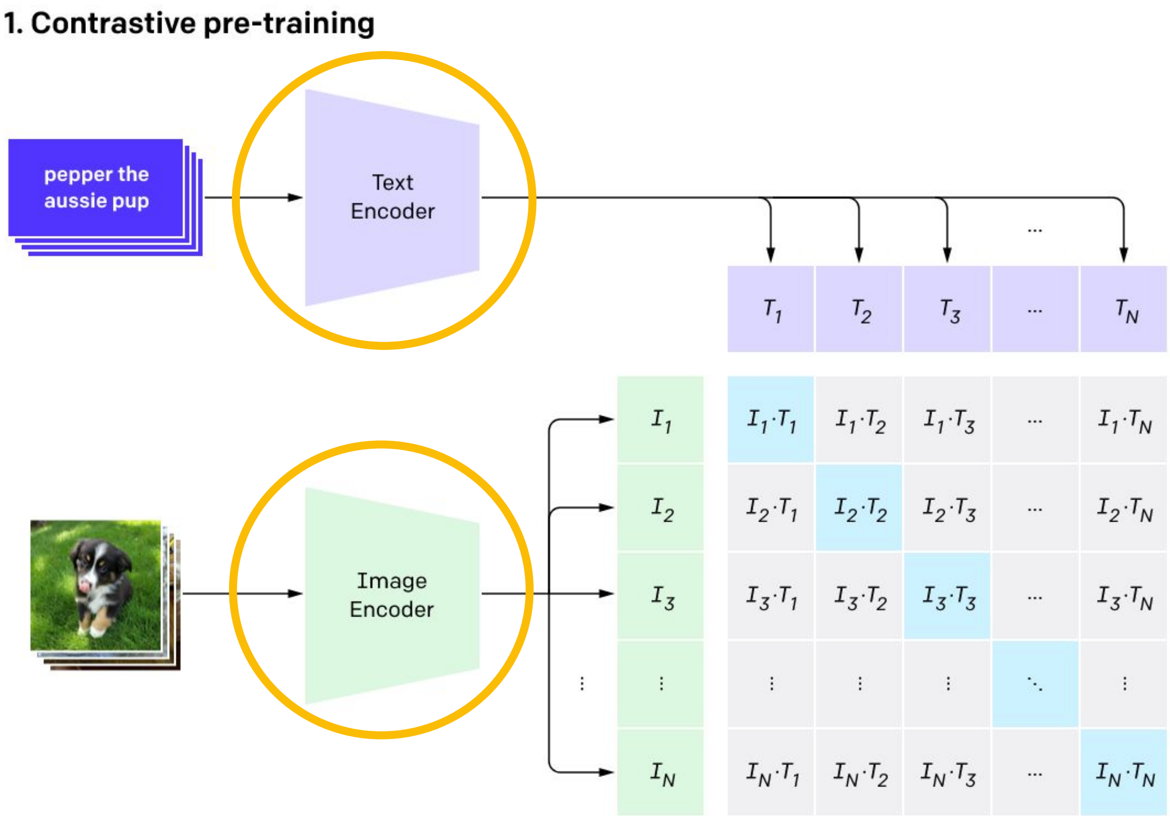
A straightforward approach:
Zero-shot detection with cropped regions

- Ensemble 1x and 1.5x crops (to include context).
- Similar to R-CNN.
- Slow!
- Not utilizing annotations from base categories.



Leveraging pre-trained zero-shot classification model

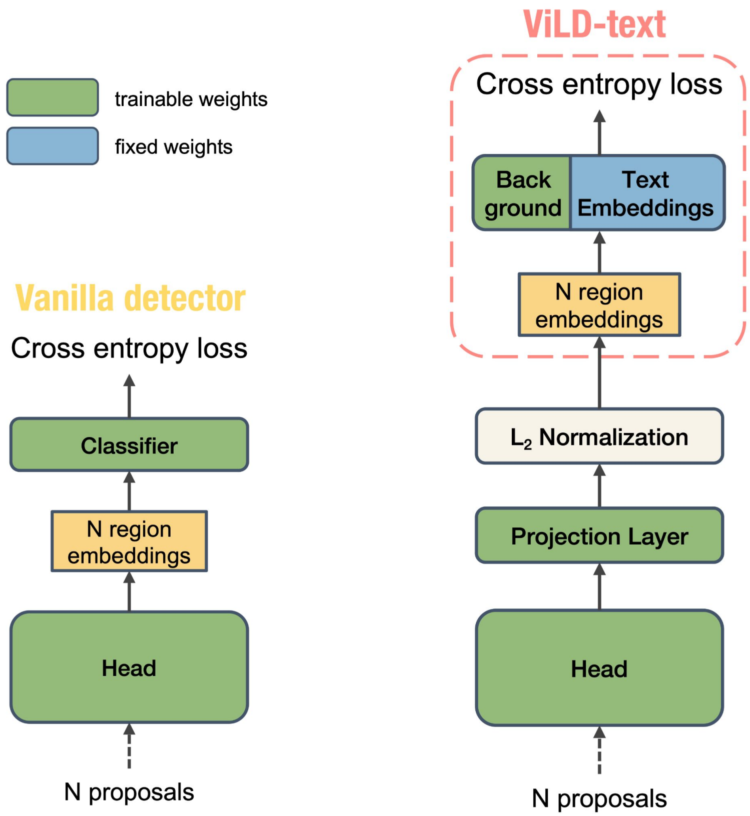
- CLIP



https://openai.com/blog/clip/

Radford et al., 2021. Learning Transferable Visual Models From Natural Language Supervision.

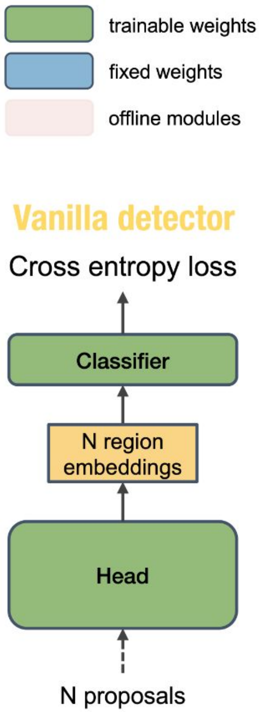
ViLD-text



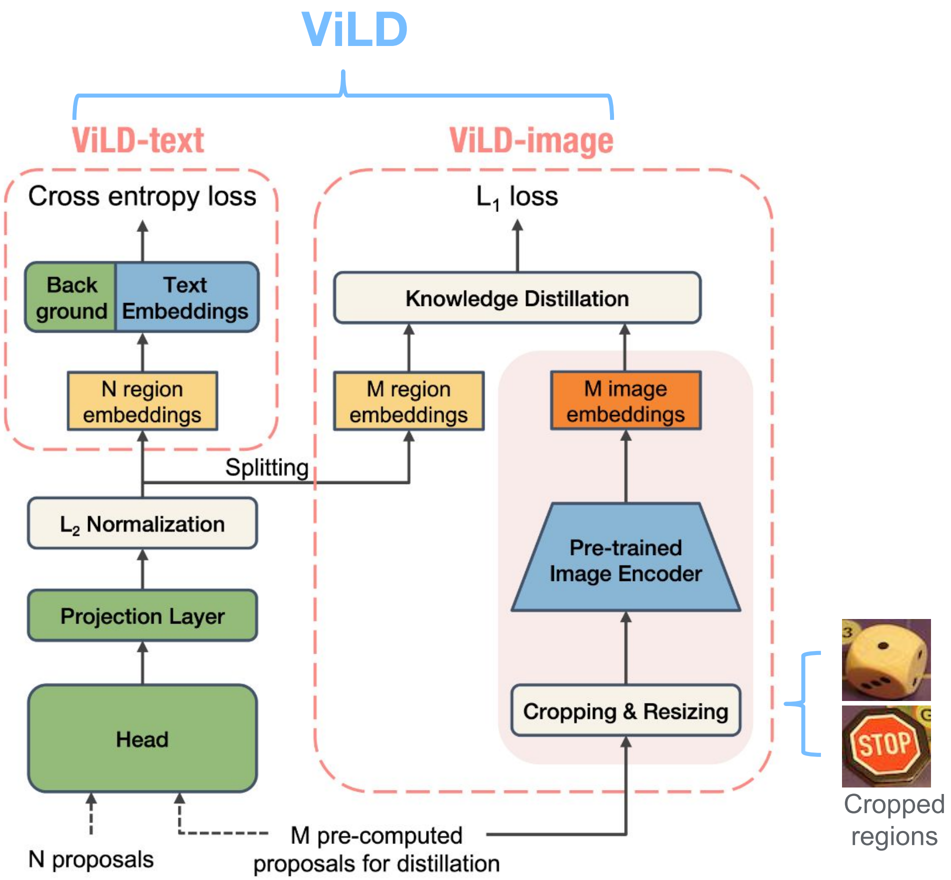
- **Text prompts**: e.g., "a photo of a {category} **in the scene**".
- **Category text embeddings**: feed the text prompts into the pre-trained text encoder.
  - Ensemble 63 text prompts, with synonyms if available.
- **Learnable "background" embedding**: for proposals do not match any labeled categories.
- **Classify with text embeddings**:

  L_{CE}(softmax(1/T * cosine_similarity(region embedding, text/background embeddings)))

ViLD-image

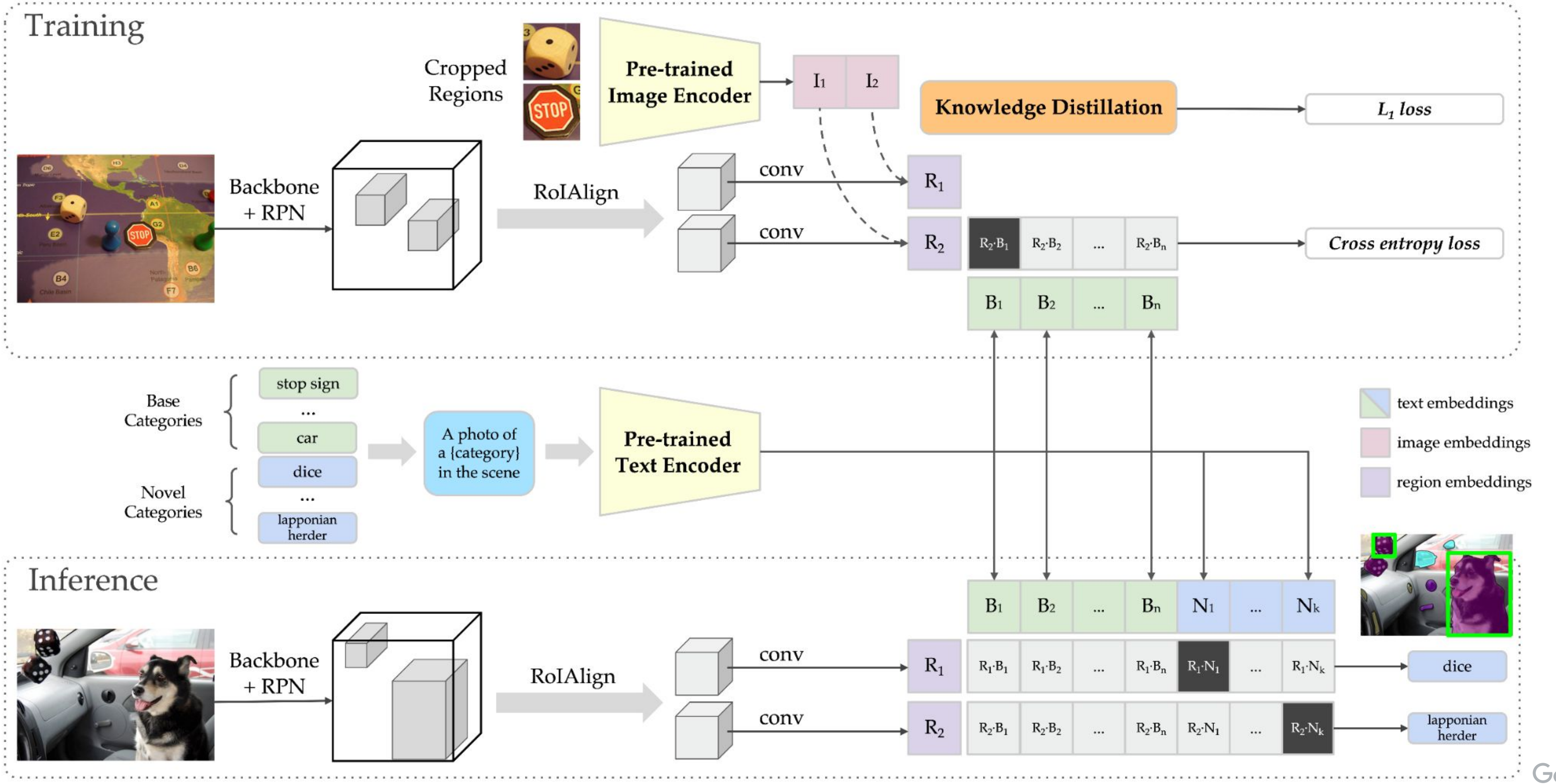


trainable weights
fixed weights
offline modules
Vanilla detector
Cross entropy loss
Classifier
N region embeddings
Head
N proposals



ViLD
ViLD-text
ViLD-image
Cross entropy loss
Back ground
Text Embeddings
N region embeddings
L_1 loss
Knowledge Distillation
M region embeddings
M image embeddings
Splitting
L_2 Normalization
Pre-trained Image Encoder
Projection Layer
Head
Cropping & Resizing
N proposals
M pre-computed proposals for distillation
STOP
Cropped regions

ViLD overview



Google

Model ensembling to mitigate conflicting objectives

- **1) ViLD-text + CLIP**:
  - CLIP is the teacher model for ViLD-image.
  - Slow!
- **2) ViLD-ensemble**: Two separate heads for ViLD-text and ViLD-image objectives, respectively.
  - **Weighted average**: For base categories, weigh the predictions of ViLD-text more; vice versa for novel categories.

Results

Benchmark settings

- Main dataset: LVIS v1.0 (1203 categories)
  - Frequent (f: 405 classes, 100-1977 images per class) and common (c: 461 classes, 10-100 images per class) categories as base categories
  - Rare (r: 337 classes, <10 images per class) categories as novel categories
- Metrics: Average Precision (AP), $\mathbf{AP_r}$, AP_c, AP_f

Object proposals for novel categories

RPN's Average Recall (AR) for novel categories

Supervision	AR_r@100	AR_r@300	AR_r@1000
base	39.3	48.3	55.6
base + novel	41.1	50.9	57.0

- RPN trained on base categories generalizes to novel categories, yielding higher scores for unseen categories compared with background.

Classifying proposals with CLIP

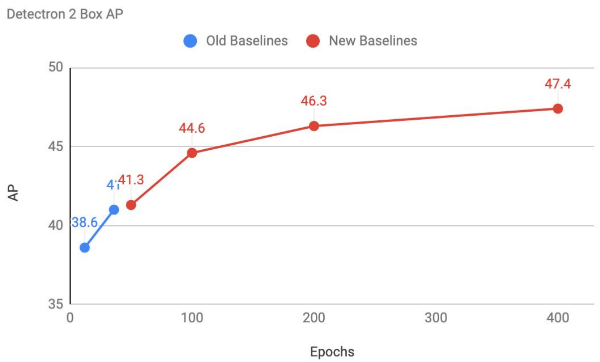
Method	AP_r	AP_c	AP_f	AP
Supervised (base class only)	0.0	22.6	32.4	22.5
CLIP on cropped regions	**13.0**	10.6	6.0	9.2
Supervised (base + novel)	4.1	23.5	33.2	23.9
Supervised-RFS (base + novel)	12.3	24.3	32.4	25.4

- Supervised-RFS: a better supervised baseline with Repeat Factor Sampling to upsample rare classes.
- AP_r is on par with supervised learning approaches.
- Overall performance is behind.

A strong baseline (large-scale jittering + longer training)

- Formally introduced in our Copy-Paste data augmentation paper [1] and recently adopted by Detectron2
- Recent FB AI Blog: [Advancing computer vision research with new Detectron2 Mask R-CNN baselines]

Recent work in the field, such as Simple Copy-Paste Data Augmentation, has shown substantial improvements in accuracy (measured by average precision, or AP) for two core tasks, creating a bounding box around an object and drawing a detailed mask over different objects. The paper's highest-reported Mask R-CNN ResNet-50-FPN baseline is 47.2 Box AP and 41.8 Mask AP, which exceeds Detectron2's highest reported baseline of 41.0 Box AP and 37.2 Mask AP. This difference is significant because most research papers publish improvements in the order of 1 percent to 3 percent.



New baselines using Large-Scale Jitter and Longer Training Schedule

The following baselines of COCO Instance Segmentation with Mask R-CNN are generated using a longer training schedule and large-scale jitter as described in Google's Simple Copy-Paste Data Augmentation paper. These models are trained from scratch using random initialization. These baselines exceed the previous Mask R-CNN baselines.

In the following table, one epoch consists of training on 118000 COCO images.

Name	epochs	train time (s/im)	inference time (s/im)	box AP	mask AP	model id	download
R50-FPN	100	0.376	0.069	44.6	40.3	42047764	model | metrics
R50-FPN	200	0.376	0.069	46.3	41.7	42047638	model | metrics
R50-FPN	400	0.376	0.069	47.4	42.5	42019571	model | metrics
R101-FPN	100	0.518	0.073	46.4	41.6	42025812	model | metrics
R101-FPN	200	0.518	0.073	48.0	43.1	42131867	model | metrics
R101-FPN	400	0.518	0.073	48.9	43.7	42073830	model | metrics

https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md

[1] Ghiasi et al., 2021. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation

ViLD-text

Backbone: Mask R-CNN R50-FPN

Method	AP_r	AP_c	AP_f	AP
CLIP on cropped regions	13.0	10.6	6.0	9.2
GloVe baseline	3.0	20.1	30.4	21.2
ViLD-text	10.1	23.9	32.5	24.9
Supervised-RFS (base + novel)	12.3	24.3	32.4	25.4

- *Outperforms GloVe baseline*:
  - text embeddings jointly trained with visual data.
- *Outperforms CLIP on cropped regions*:
  - trained with annotations from base categories.

ViLD-image

Backbone: Mask R-CNN R50-FPN

Method	AP_r	AP_c	AP_f	AP
CLIP on cropped regions	13.0	10.6	6.0	9.2
GloVe baseline	3.0	20.1	30.4	21.2
ViLD-text	10.1	23.9	32.5	24.9
ViLD-image	9.6	8.5	7.8	8.4
Supervised-RFS (base + novel)	12.3	24.3	32.4	25.4

- Only trained with L1 distillation loss, no cross entropy loss.
- Ideally, similar performance as CLIP on cropped regions.
- A small performance gap.

ViLD

Backbone: Mask R-CNN R50-FPN

Method	AP_r	AP_c	AP_f	AP
CLIP on cropped regions	13.0	10.6	6.0	9.2
GloVe baseline	3.0	20.1	30.4	21.2
ViLD-text	10.1	23.9	32.5	24.9
ViLD-image	9.6	8.5	7.8	8.4
ViLD ($w = 0.5$)	16.1	20.0	28.3	22.5
Supervised-RFS (base + novel)	12.3	24.3	32.4	25.4

- **Outperforms supervised counterpart on novel categories!**
- AP_r improved over ViLD-text or ViLD-image.

ViLD

Hyperparameter sweep

Distill loss	Distill weight w	AP_r	AP_c	AP_f	AP
No distill	0.0	10.4	22.9	31.3	24.0
$\mathcal{L}_2$ loss	0.5	**13.7**	21.7	31.2	24.0
	1.0	12.4	22.7	31.4	24.3
	2.0	13.4	22.0	30.9	24.0
$\mathcal{L}_1$ loss	0.05	12.9	22.4	31.7	24.4
	0.1	14.0	20.9	31.2	23.8
	0.5	16.3	19.2	27.3	21.9
	1.0	**17.3**	18.2	25.1	20.7

- L1 loss is better than L2 loss.
- Trend: as *w* increases, APr ↑, APf and APc ↓ →

  A competition between ViLD-text and ViLD-image.
- We later mitigate the competition by ensembling.

Model ensembling

Backbone: Mask R-CNN R50-FPN

Method	AP_r	AP_c	AP_f	AP
CLIP on cropped regions	13.0	10.6	6.0	9.2
GloVe baseline	3.0	20.1	30.4	21.2
ViLD-text	10.1	23.9	32.5	24.9
ViLD-image	9.6	8.5	7.8	8.4
ViLD ($w = 0.5$)	16.1	20.0	28.3	22.5
ViLD-ensemble ($w = 0.5$)	**16.6**	24.6	30.3	25.5
ViLD-text + CLIP †	**22.6**	24.8	29.2	26.1
Supervised-RFS (base + novel)	12.3	24.3	32.4	25.4

- ViLD-text + CLIP attains the best AP_r and good overall AP.
  - 630x slower.
- ViLD-ensemble improves AP_c and AP_f over ViLD.

Results with Mask R-CNN R152-FPN backbone

Backbone: Mask R-CNN R152-FPN

Method	AP_r	AP_c	AP_f	AP
ViLD-text	11.7	25.8	34.4	26.7
ViLD-image	10.8	10.0	8.7	9.6
ViLD ($w = 1.0$)	**18.7**	21.1	28.4	23.6
ViLD-ensemble ($w = 2.0$)	**18.7**	24.9	30.6	26.0
Supervised-RFS (base + novel)	14.4	26.8	34.2	27.6

- AP_r further improves with stronger backbones.
- Same trend as R50.

Transfer to other detection datasets

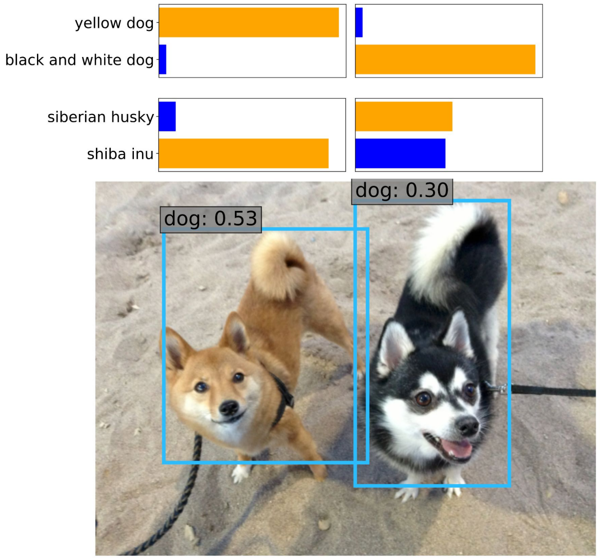
- Replace with category text embeddings of a new dataset.
- **A finetuning-free transfer!**
- Small gaps compared with finetuning (start from ViLD, finetune the linear classifier).

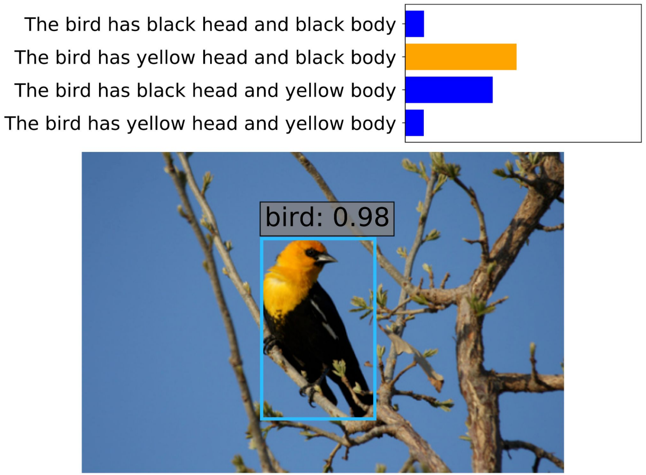
Method	PASCAL VOC†		COCO						Objects365					
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ViLD-text	40.5	31.6	28.8	43.4	31.4	11.8	35.5	52.0	10.4	15.8	11.1	4.5	11.4	20.4
ViLD	72.2	56.7	36.6	55.6	39.8	20.7	39.2	52.6	11.8	18.2	12.6	5.5	13.5	21.2
Finetuning	78.9	60.3	39.1	59.8	42.4	21.0	41.7	55.0	15.2	23.9	16.2	7.3	17.2	26.1
Supervised	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

Qualitative examples

On-the-fly interactive detection

- After detecting pre-defined categories, use *on-the-fly free-form* text embeddings to recognize more details.





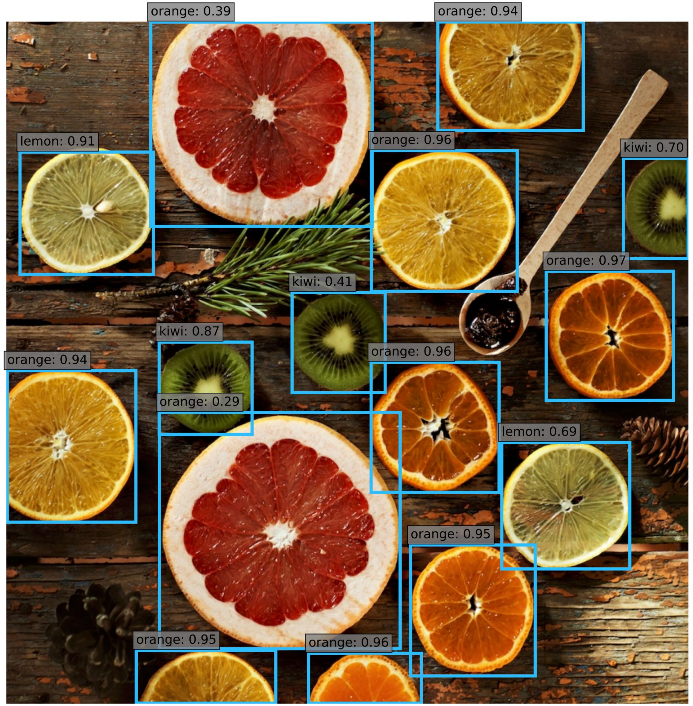
Systematic expansion of dataset vocabulary

- Dataset vocabulary: $\mathbf{v} = \{v_1, \ldots, v_p\}$.
- Attributes set: $\mathbf{a} = \{a_1, \ldots, a_q\}$.
- Given region embedding $\mathbf{e}_r$,

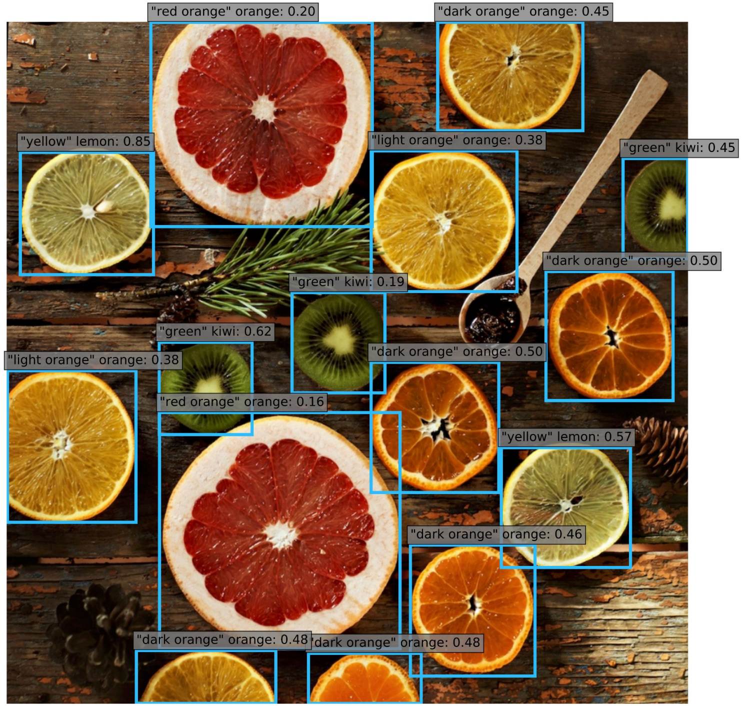
  $\Pr(v_i, a_j \mid \mathbf{e}_r) = \Pr(v_i \mid \mathbf{e}_r) * \Pr(a_j \mid \mathbf{e}_r)$.
- Expand p vocabularies into $p * q$ vocabularies.

Systematic expansion of dataset vocabulary

- Detect fruit with color attributes (expand LVIS vocabulary with 11 colors).



Original dataset vocabulary

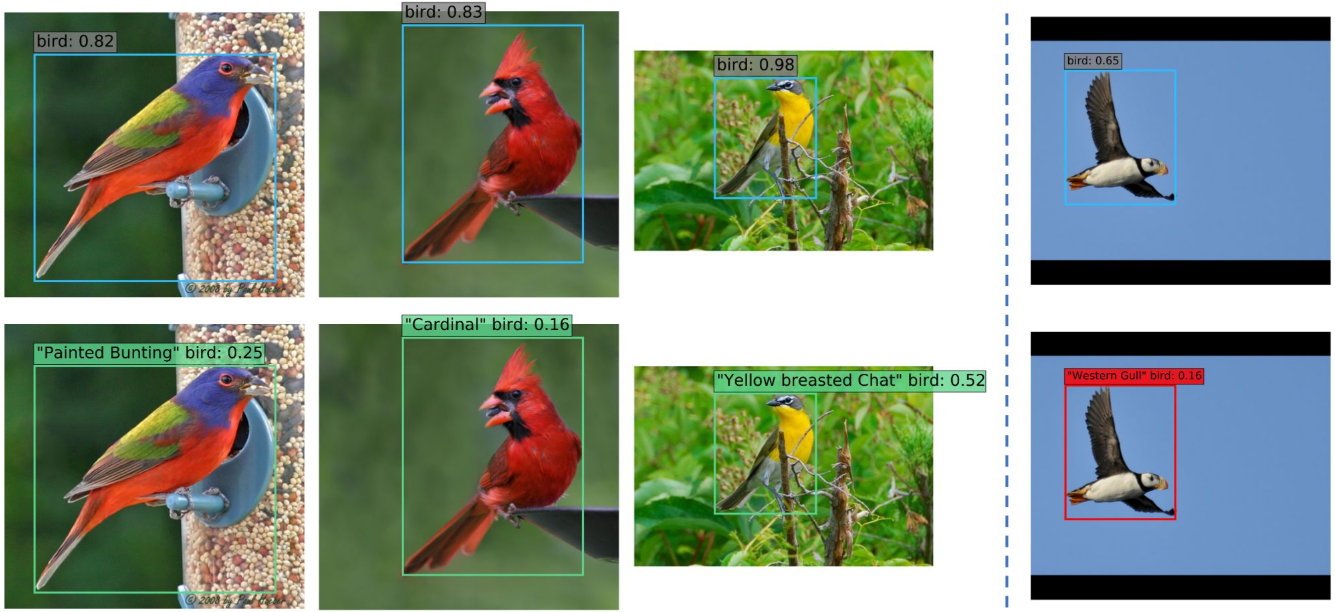


Expanded with color attributes

Google

Systematic expansion of dataset vocabulary

- 200 Fine-grained bird species from CUB-200-2011 (expanded from LVIS vocabulary).

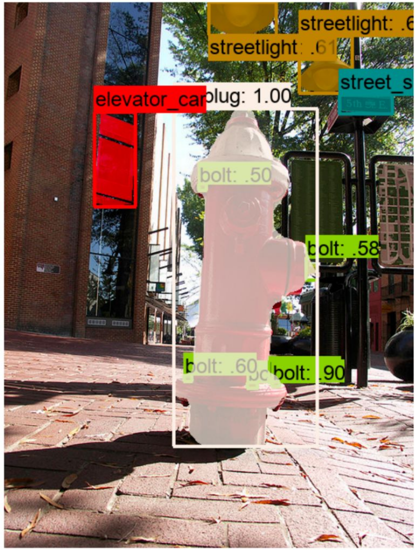


(a) Successful cases

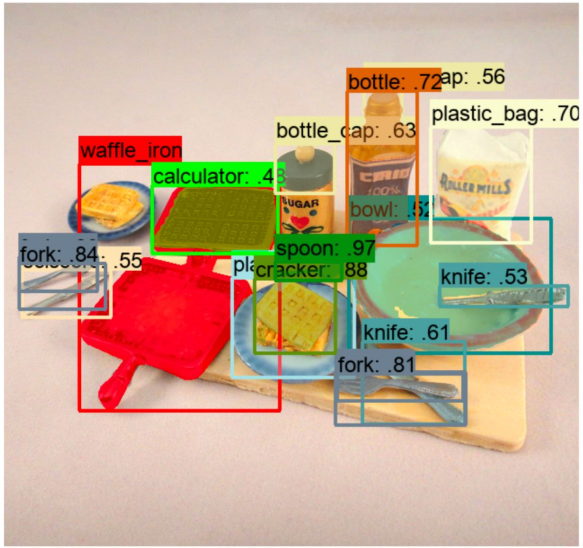
(b) Failure case

Failure cases

- red: groundtruth of failed detections for novel objects



(a) Missed



(b) Misclassified

Compare w/ existing zero-shot detection methods on COCO

- An issue of zero-shot detention on COCO: people are using different splits of COCO dataset
- ViLD-text alone outperforms the most recent SOTA under the same setting (65/15 split, IoU=0.5), especially Recall

Method	Unseen (mAP/Recall@100)	Seen (mAP/Recall@100)
Rahman *et al.*	12.40 / 37.16	34.07 / 36.38
ViLD-text	**14.71 / 47.69**	**43.62 / 85.55**

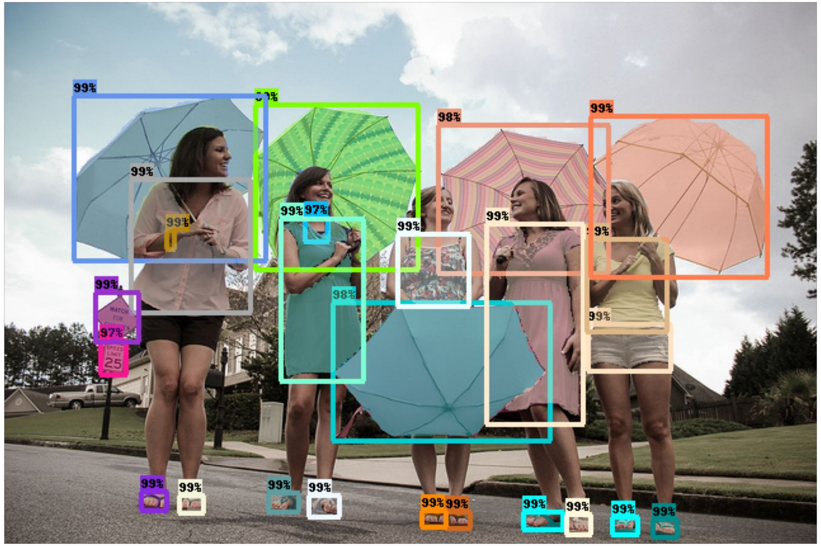
Rahman et al., 2020. Improved visual-semantic alignment for zero-shot object detection

An improved version with ALIGN

- ViLD-ensemble on LVIS v1.0 with a pre-trained ALIGN model

Method	Image Model for Distillation	Detector Backbone	APr (novel)	APc (base)	APf (base)
ViLD-ensemble	CLIP (ViT-B/32)	ResNet-152	18.7	24.9	30.6
ViLD-ensemble	ALIGN (EfficientNet-B7)	EfficientNet-B7	26.3	27.2	32.9
Supervised Baseline	-	EfficientNet-B7	15.4 (-10.9)	27.8 (+0.6)	34.3 (+1.4)

An improved version with ALIGN - qualitative example



Input texts

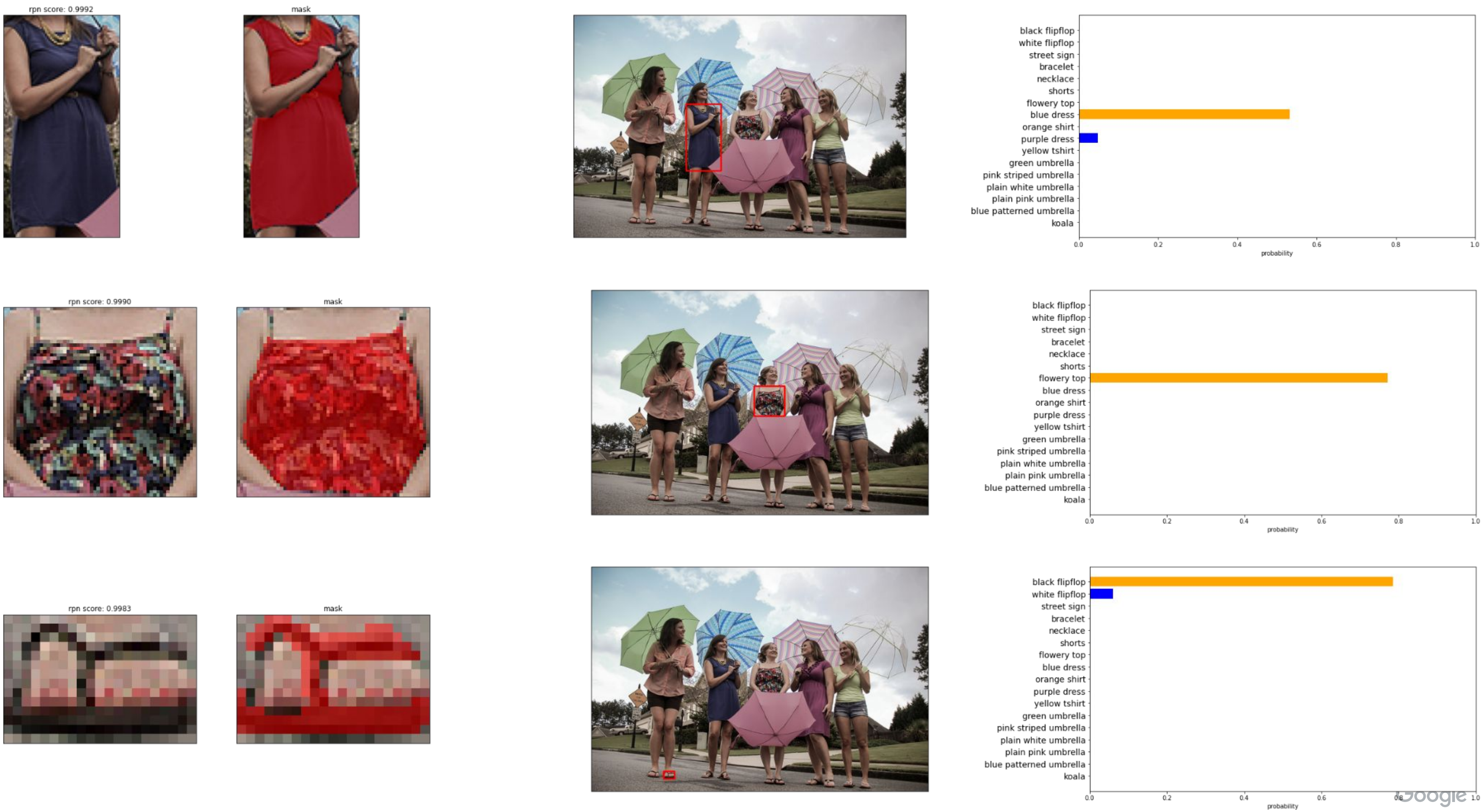
```
category_indices:
{1: {'id': 1, 'name': 'black flipflop'},
 2: {'id': 2, 'name': 'white flipflop'},
 3: {'id': 3, 'name': 'street sign'},
 4: {'id': 4, 'name': 'bracelet'},
 5: {'id': 5, 'name': 'necklace'},
 6: {'id': 6, 'name': 'shorts'},
 7: {'id': 7, 'name': 'flowery top'},
 8: {'id': 8, 'name': 'blue dress'},
 9: {'id': 9, 'name': 'orange shirt'},
 10: {'id': 10, 'name': 'purple dress'},
 11: {'id': 11, 'name': 'yellow tshirt'},
 12: {'id': 12, 'name': 'green umbrella'},
 13: {'id': 13, 'name': 'pink striped umbrella'},
 14: {'id': 14, 'name': 'plain white umbrella'},
 15: {'id': 15, 'name': 'plain pink umbrella'},
 16: {'id': 16, 'name': 'blue patterned umbrella'},
 17: {'id': 17, 'name': 'koala'}}
```


rpn score: 0.9996
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability


rpn score: 0.9994
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

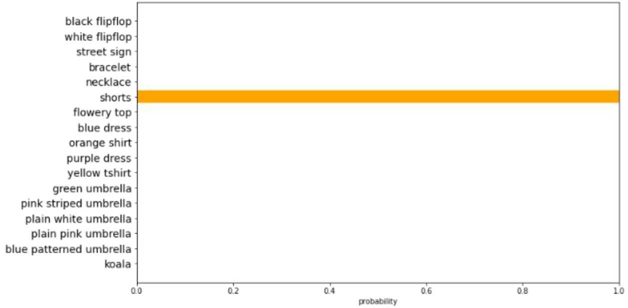

rpn score: 0.9992
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

Google


rpn score: 0.9992
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability
rpn score: 0.9990
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability
rpn score: 0.9983
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

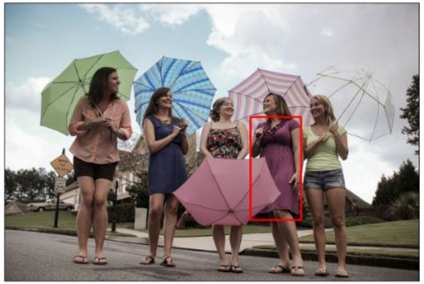
rpn score: 0.9979

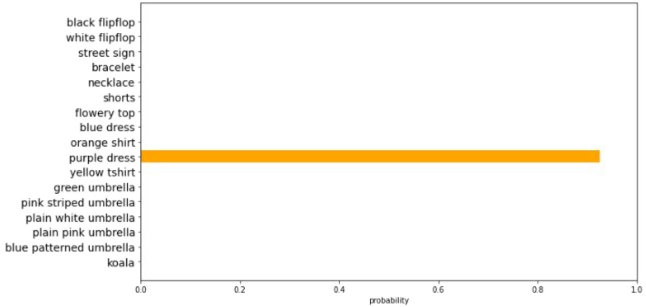
mask



rpn score: 0.9976

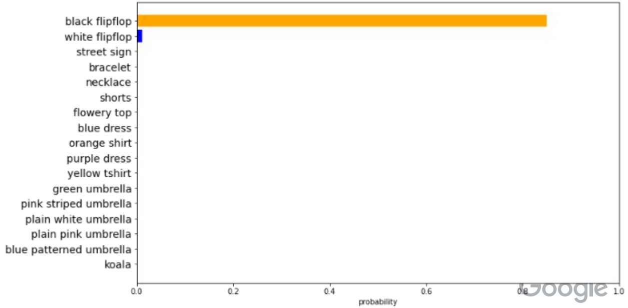
mask





rpn score: 0.9975

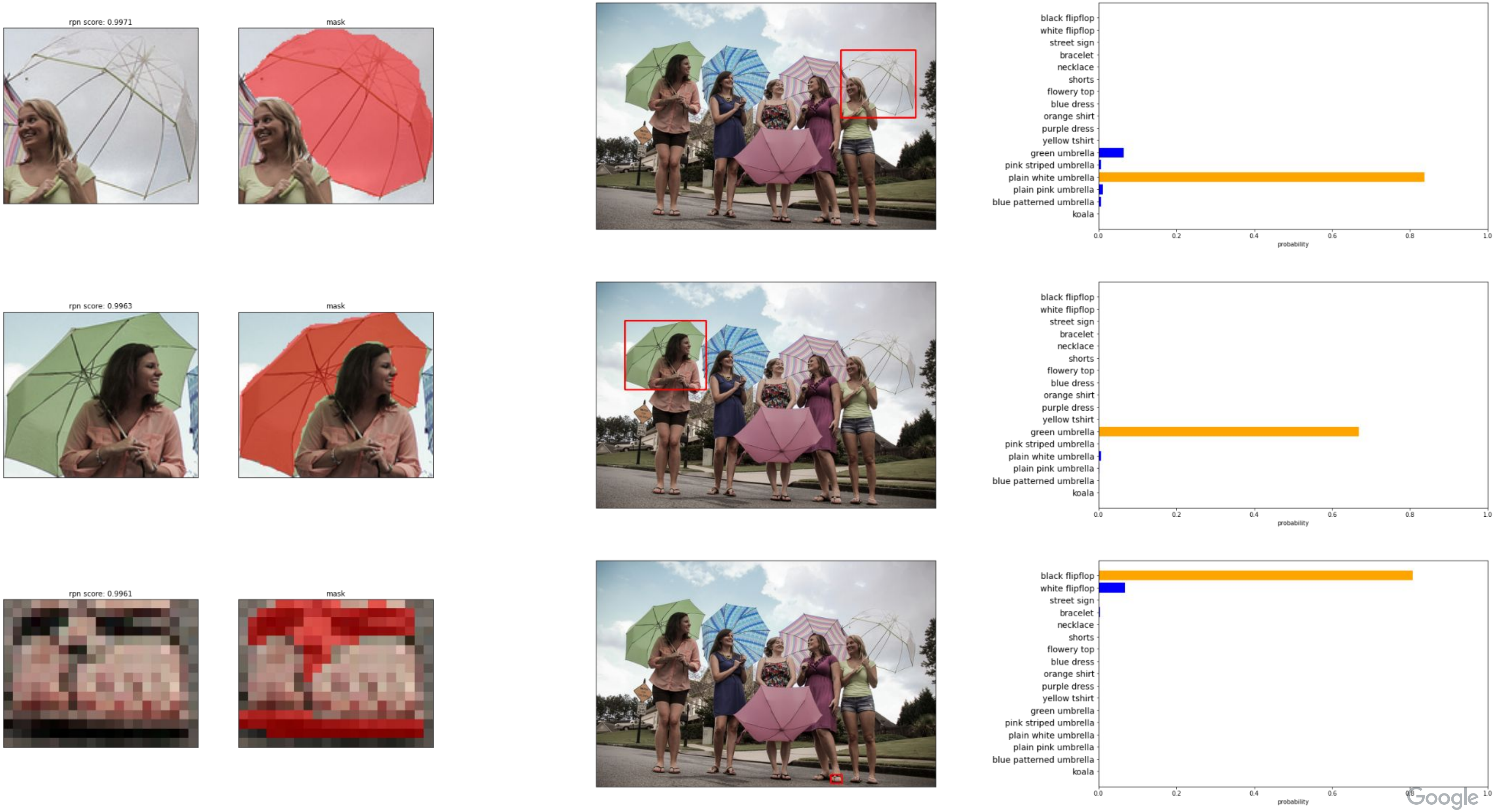
mask



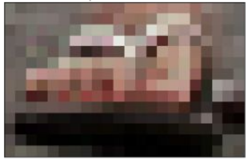

rpn score: 0.9974
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability


rpn score: 0.9974
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

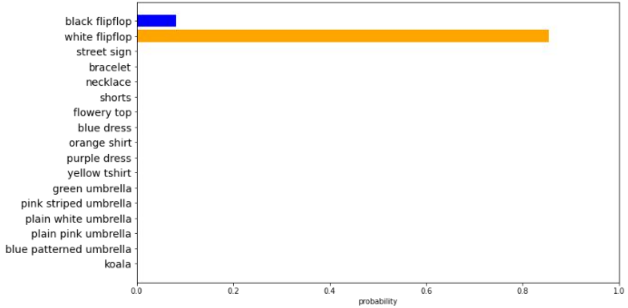

rpn score: 0.9971
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability


rpn score: 0.9971
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0 0.2 0.4 0.6 0.8 1.0
probability
rpn score: 0.9963
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0 0.2 0.4 0.6 0.8 1.0
probability
rpn score: 0.9961
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0 0.2 0.4 0.6 0.8 1.0
probability
Google

rpn score: 0.9951

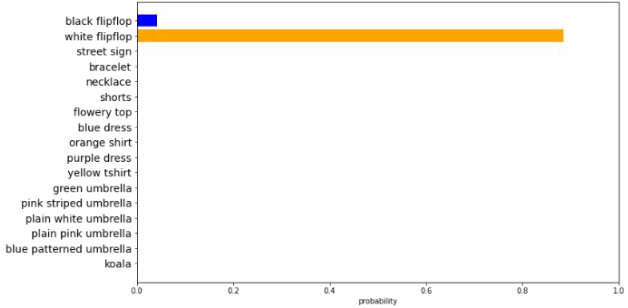


mask



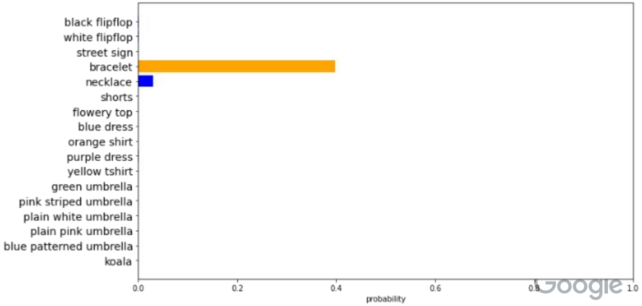
rpn score: 0.9949

mask



rpn score: 0.9932

mask



rpn score: 0.9926
mask
rpn score: 0.9917
mask
rpn score: 0.9911
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

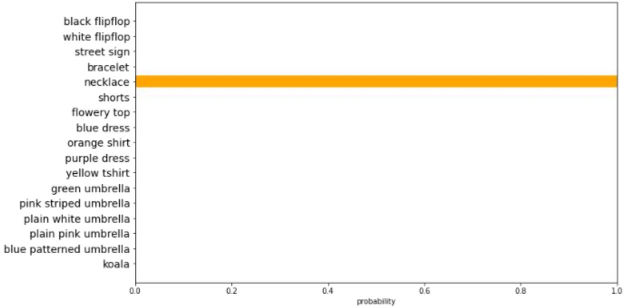

rpn score: 0.9885
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability


rpn score: 0.9870
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability


rpn score: 0.9747
mask
black flipflop
white flipflop
street sign
bracelet
necklace
shorts
flowery top
blue dress
orange shirt
purple dress
yellow tshirt
green umbrella
pink striped umbrella
plain white umbrella
plain pink umbrella
blue patterned umbrella
koala
0.0
0.2
0.4
0.6
0.8
1.0
probability

rpn score: 0.9724

mask



Conclusion

- ViLD: an open-vocabulary detection method by distilling knowledge from a zero-shot image classification model.
- Achieves 18.7 AP_r on LVIS (R152-FPN), surpassing supervised counterpart at the same inference speed.
- Append new classes without re-training of the detector.
- Transfers to other datasets without fine-tuning.
- Enables free-form text detection.
- An alternative for detecting long-tailed classes, rather than scaling up detection datasets by collecting exponentially more images to cover long-tail classes.

Thank You

Google