PromptDet: Towards Open-vocabulary Detection using Uncurated Images

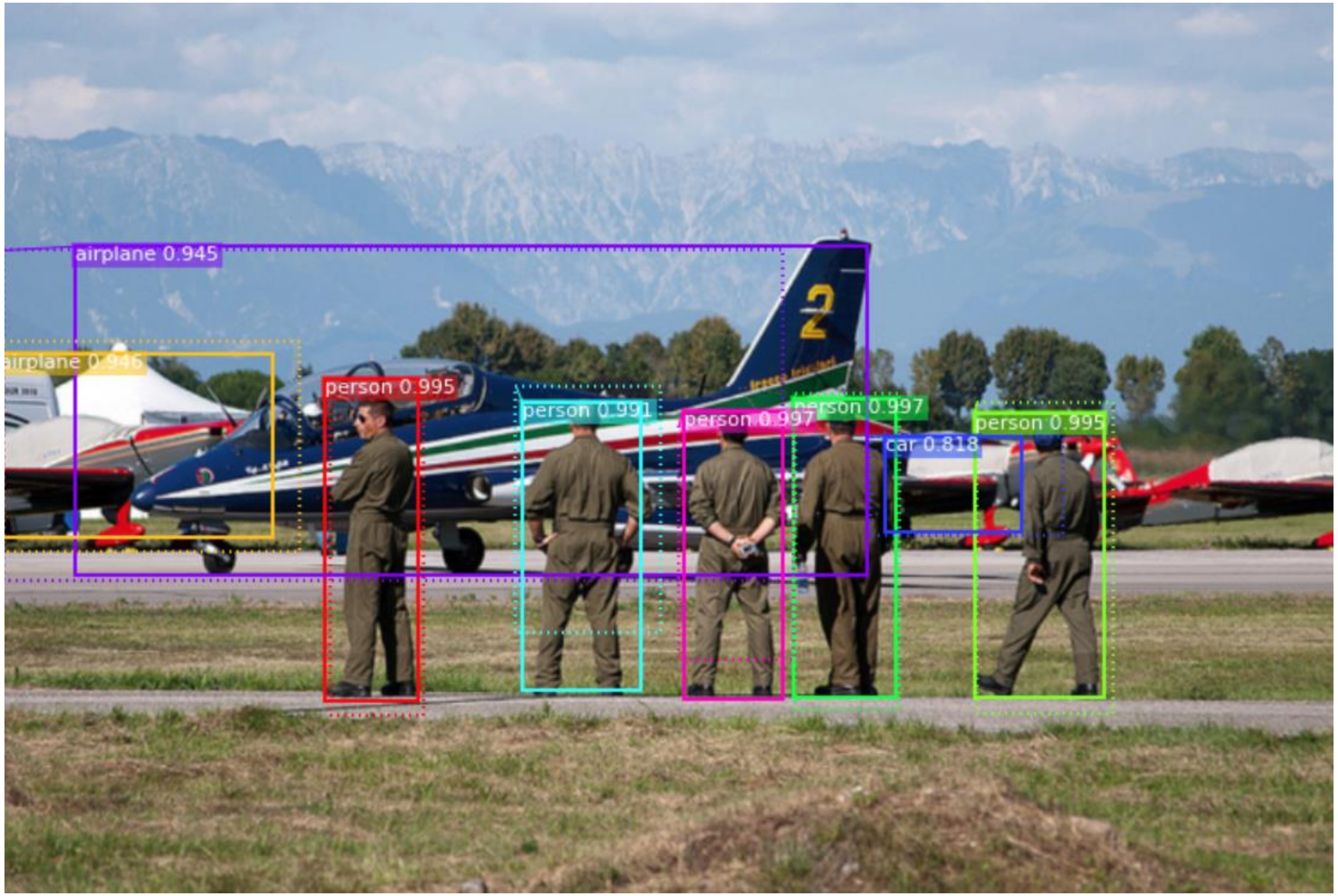
Chengjian Feng[1], Yujie Zhong[1], Zequn Jie[1], Xiangxiang Chu[1]
Haibing Ren[1], Xiaolin Wei[1], Weidi Xie[2,✉], and Lin Ma[1]

[1]Meituan Inc. [2]Shanghai Jiao Tong University

韩坤洋

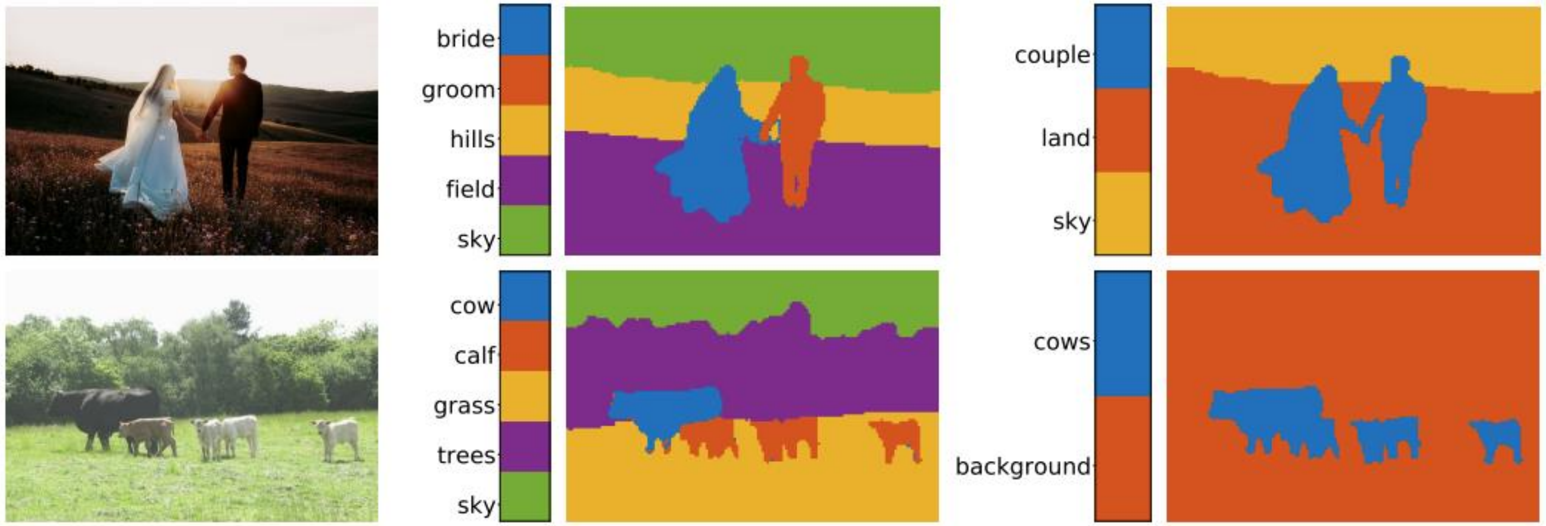
Detection

- Localization
- Classification

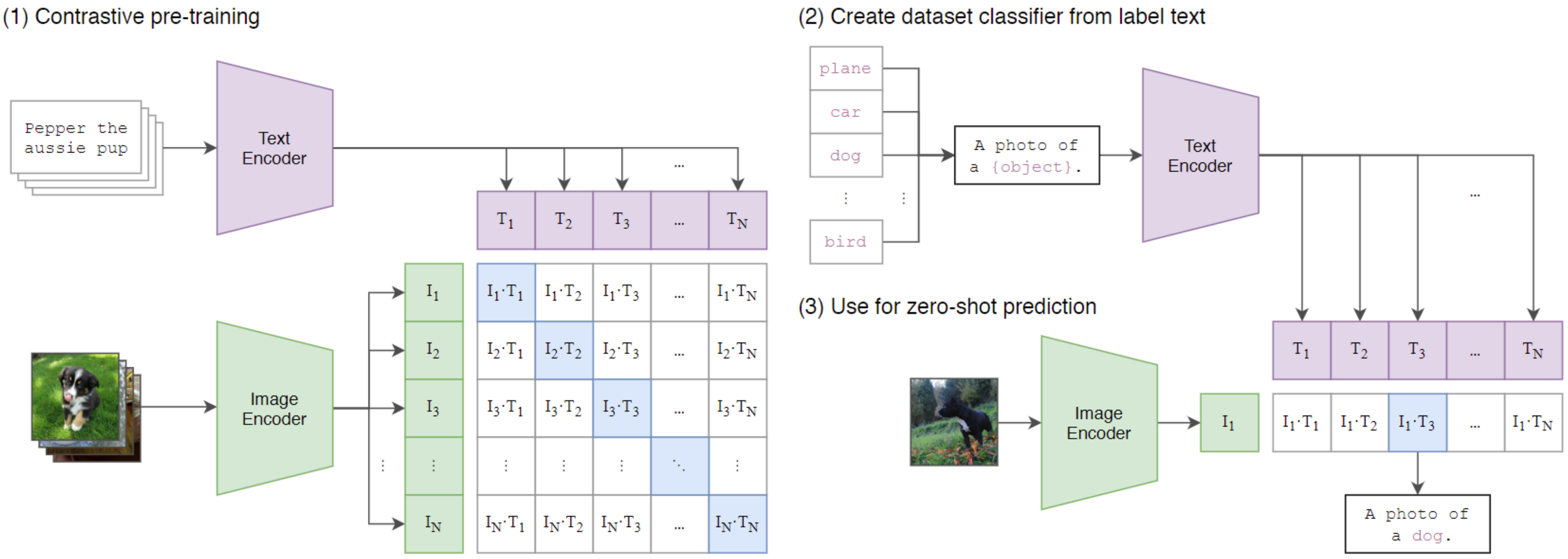


Open-vocabulary

- Open-set
  - Train on A, test on B
- Multimodality
  - Visual, image/video
  - Language, text
- Prediction related with input text prompt

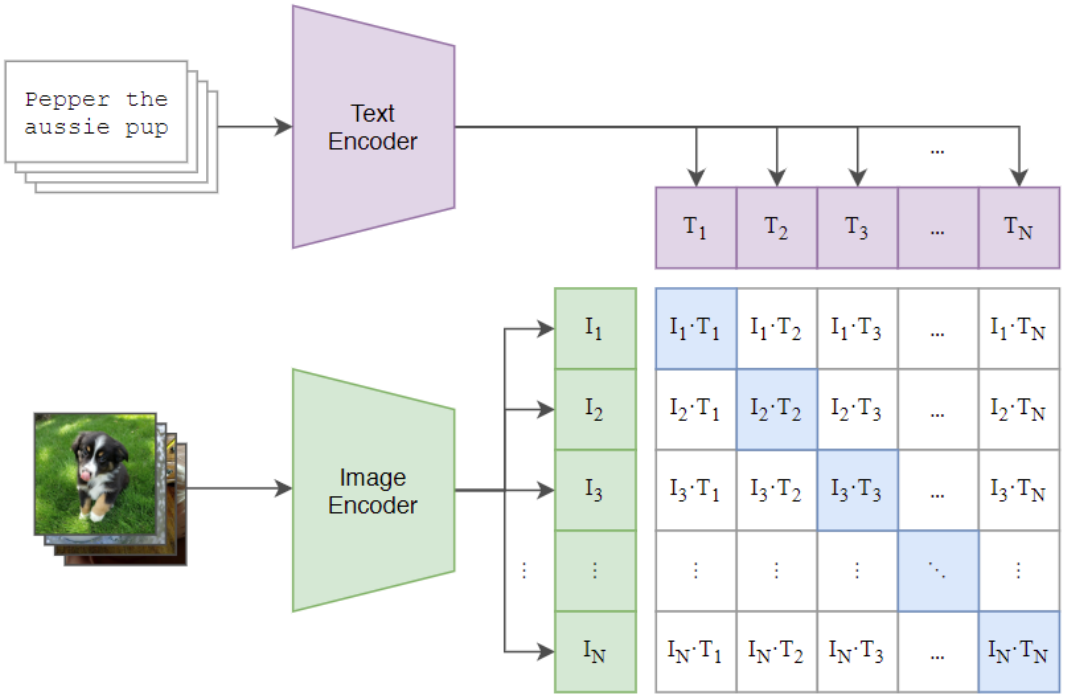


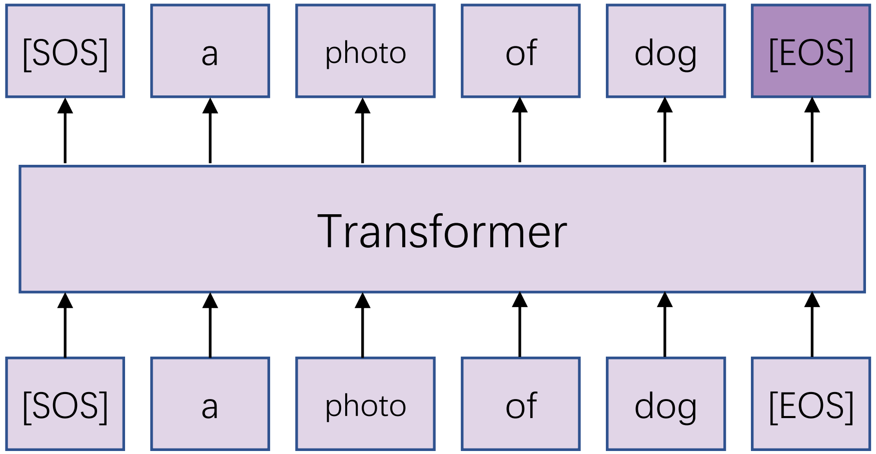
CLIP: Connecting Text and Images



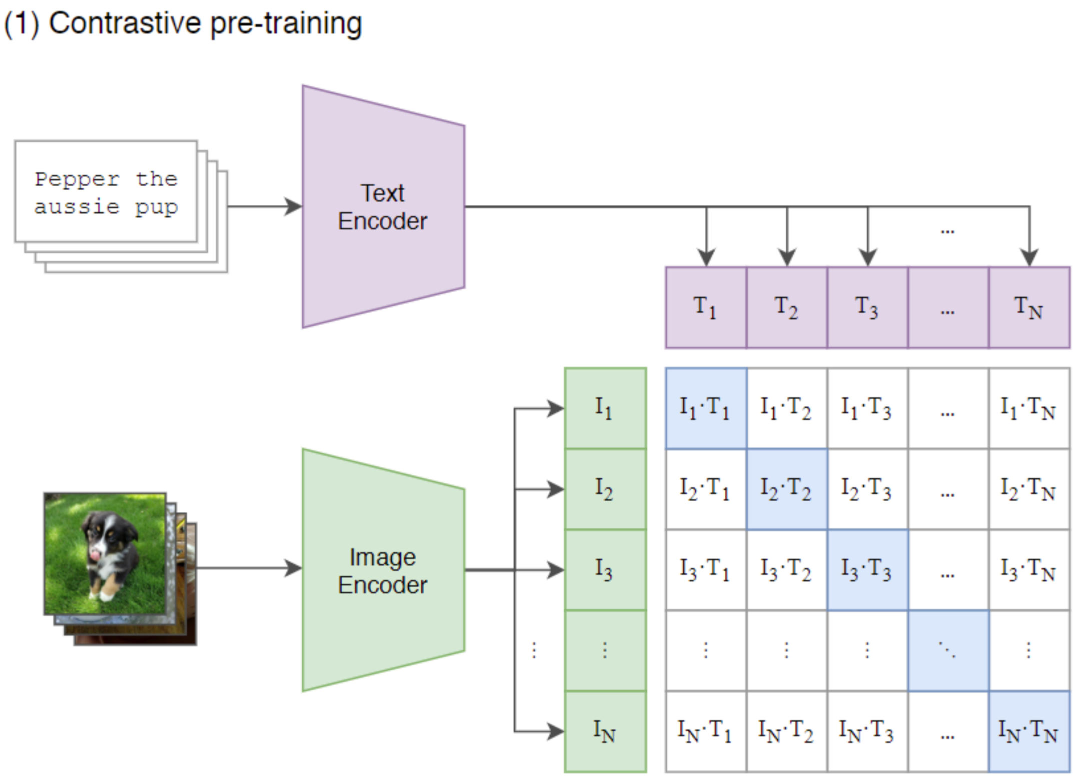
CLIP: Connecting Text and Images

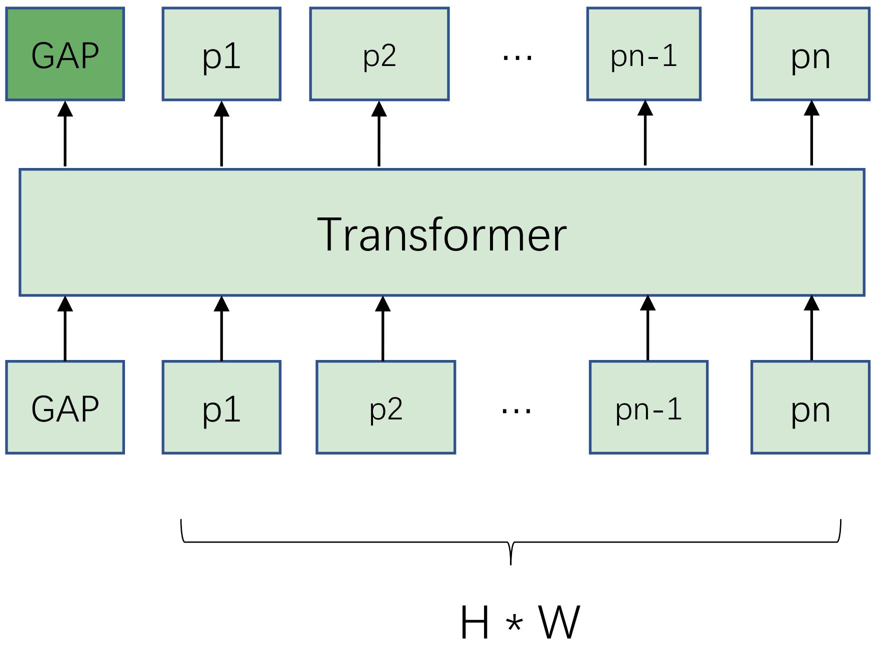
(1) Contrastive pre-training



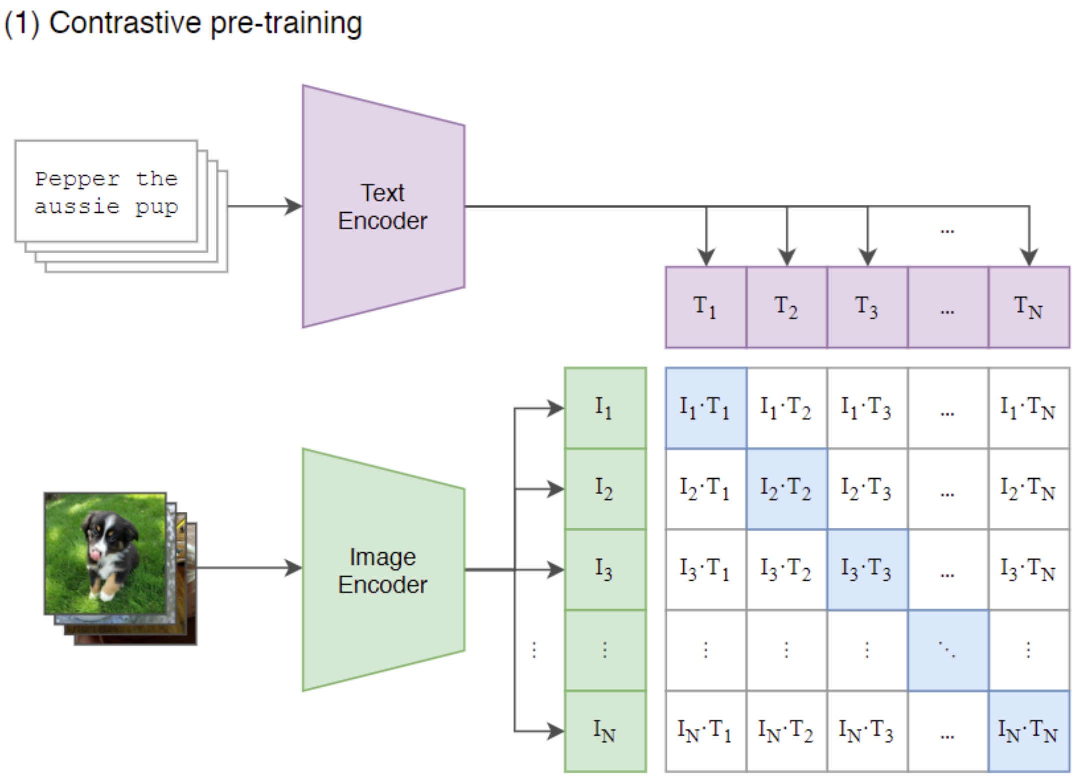


CLIP: Connecting Text and Images





CLIP: Connecting Text and Images




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

CLIP: Connecting Text and Images

- Dataset
  - 400 million pairs (image, text)
  - Collected from Internet
- Training
  - 32 epochs
  - minibatch size of 32,768
  - 18 days on 592 V100 GPUs (RN50x64)
  - 12 days on 256 V100 GPUs (ViT-L/14)

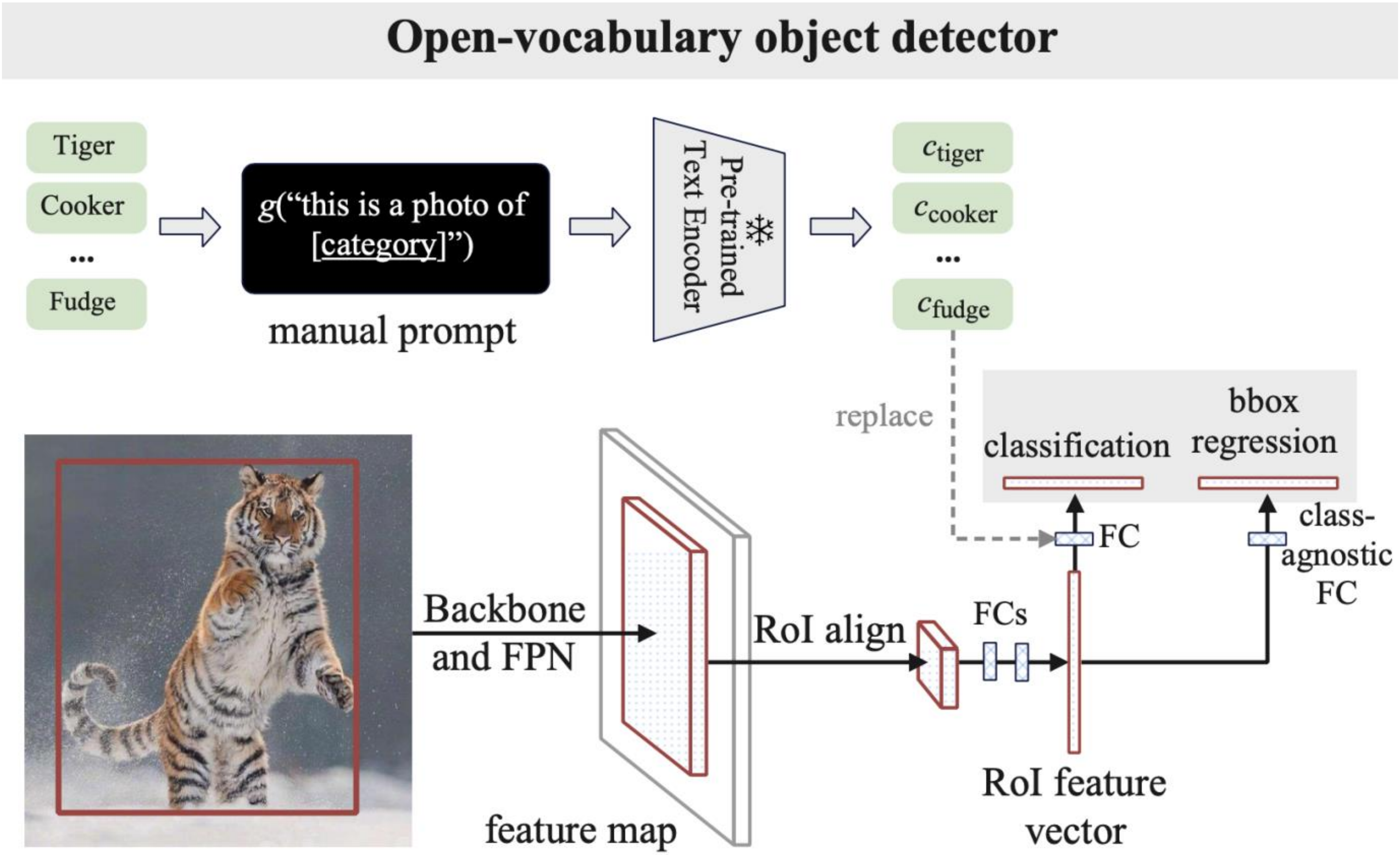
PromptDet: Towards Open-vocabulary Detection using Uncurated Images

Chengjian Feng[1], Yujie Zhong[1], Zequn Jie[1], Xiangxiang Chu[1]
Haibing Ren[1], Xiaolin Wei[1], Weidi Xie[2,✉], and Lin Ma[1]

[1]Meituan Inc. [2]Shanghai Jiao Tong University

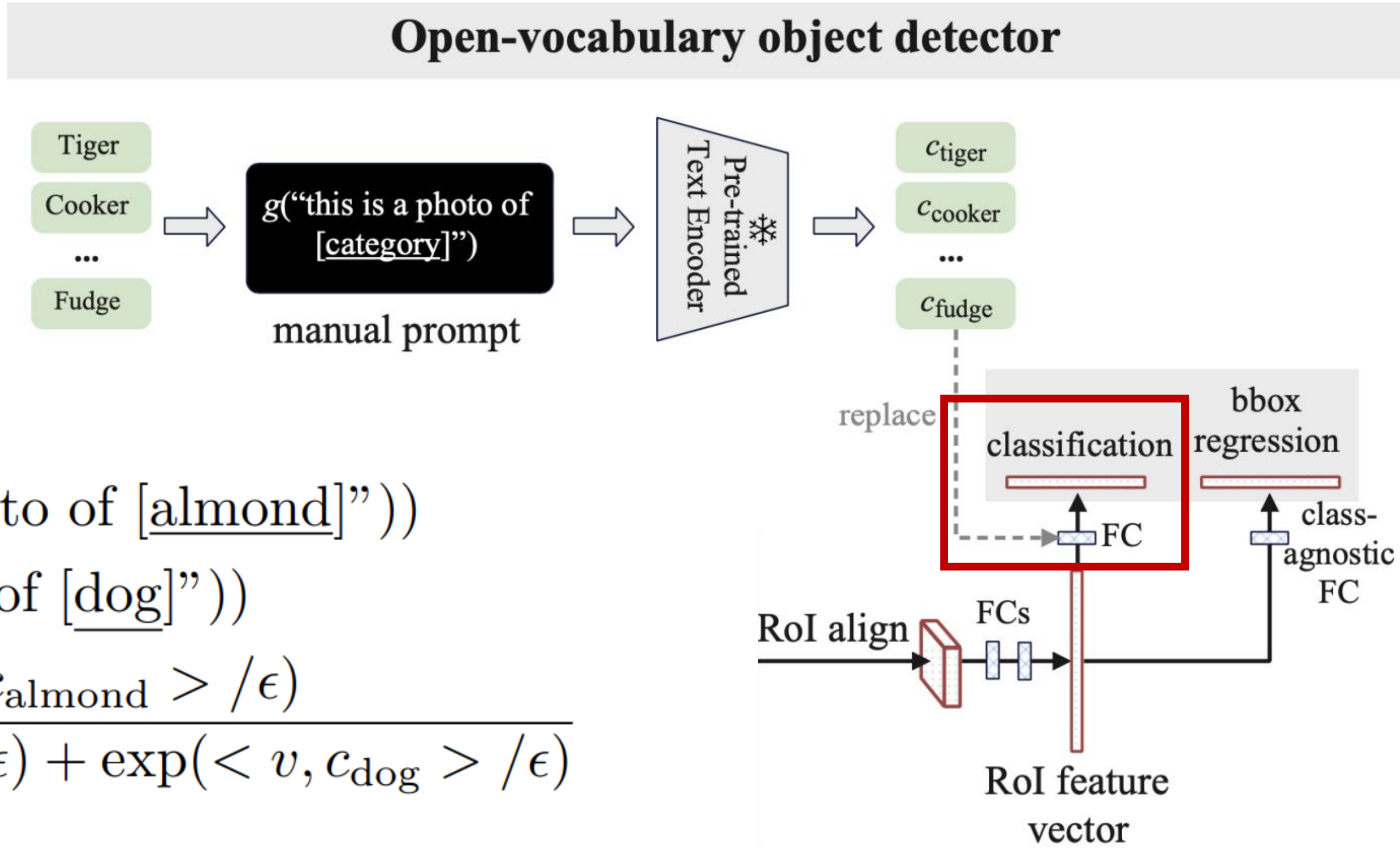
Naive Detector

- Class-agnostic RPN
- Classify with distance



Naive Detector

- Class-agnostic RPN
- Classify with distance



$$c_{\text{almond}} = \phi_{\text{text}}(g(\text{“this is a photo of }[\underline{\text{almond}}]\text{”}))$$

$$c_{\text{dog}} = \phi_{\text{text}}(g(\text{“this is a photo of }[\underline{\text{dog}}]\text{”}))$$

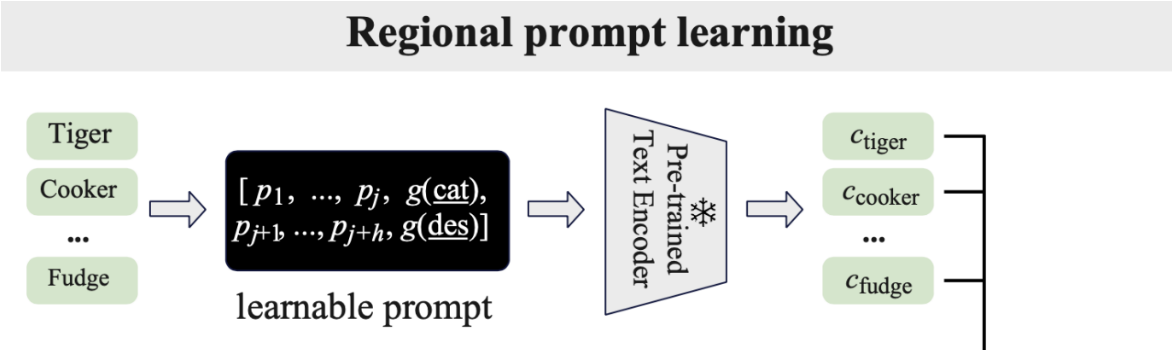
$$p_{\text{almond}} = \frac{\exp(< v, c_{\text{almond}} > /\epsilon)}{\exp(< v, c_{\text{almond}} > /\epsilon) + \exp(< v, c_{\text{dog}} > /\epsilon)}$$

Limitation

- {Class name} → category embedding, suboptimal
    - Lexical ambiguity
- Domain gap on the visual representation
    - CLIP, scene-centric
    - RPN, object-centric
- Base categories, less diverse
    - Insufficient to guarantee the generalization

Regional Prompt Learning RPL

- Learnable vectors
  - Shared for all categories
- Description

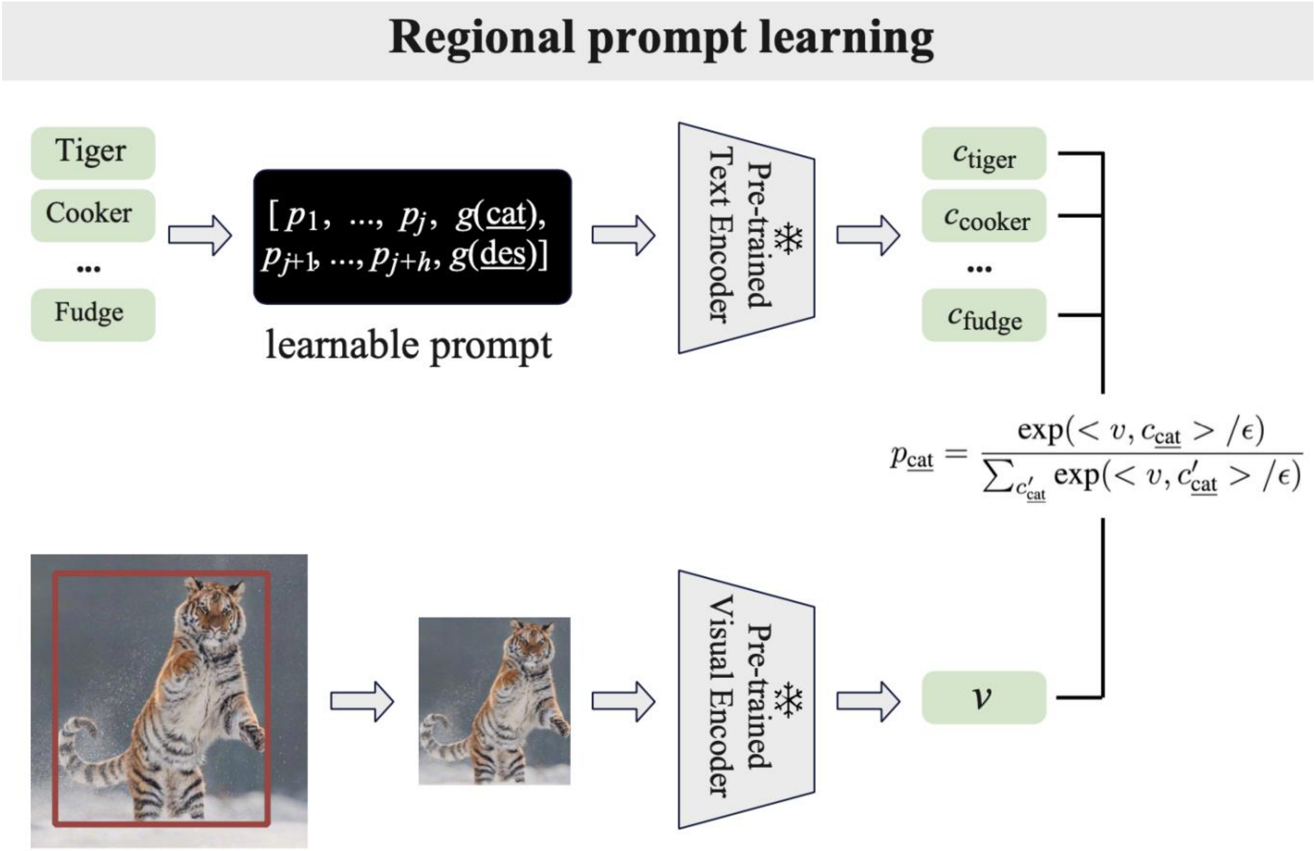


$$c_{\text{almond}} = \phi_{\text{text}}([p_1, \ldots, p_j, g(\underline{\text{category}}), p_{j+1} \ldots, p_{j+h}, g(\underline{\text{description}})])$$

{category: “almond”, description: “oval-shaped edible seed of the almond tree”}

Regional Prompt Learning RPL

- Off-line manner
- Visual
    - Crops from LVIS
    - Base categories



Regional Prompt Learning RPL

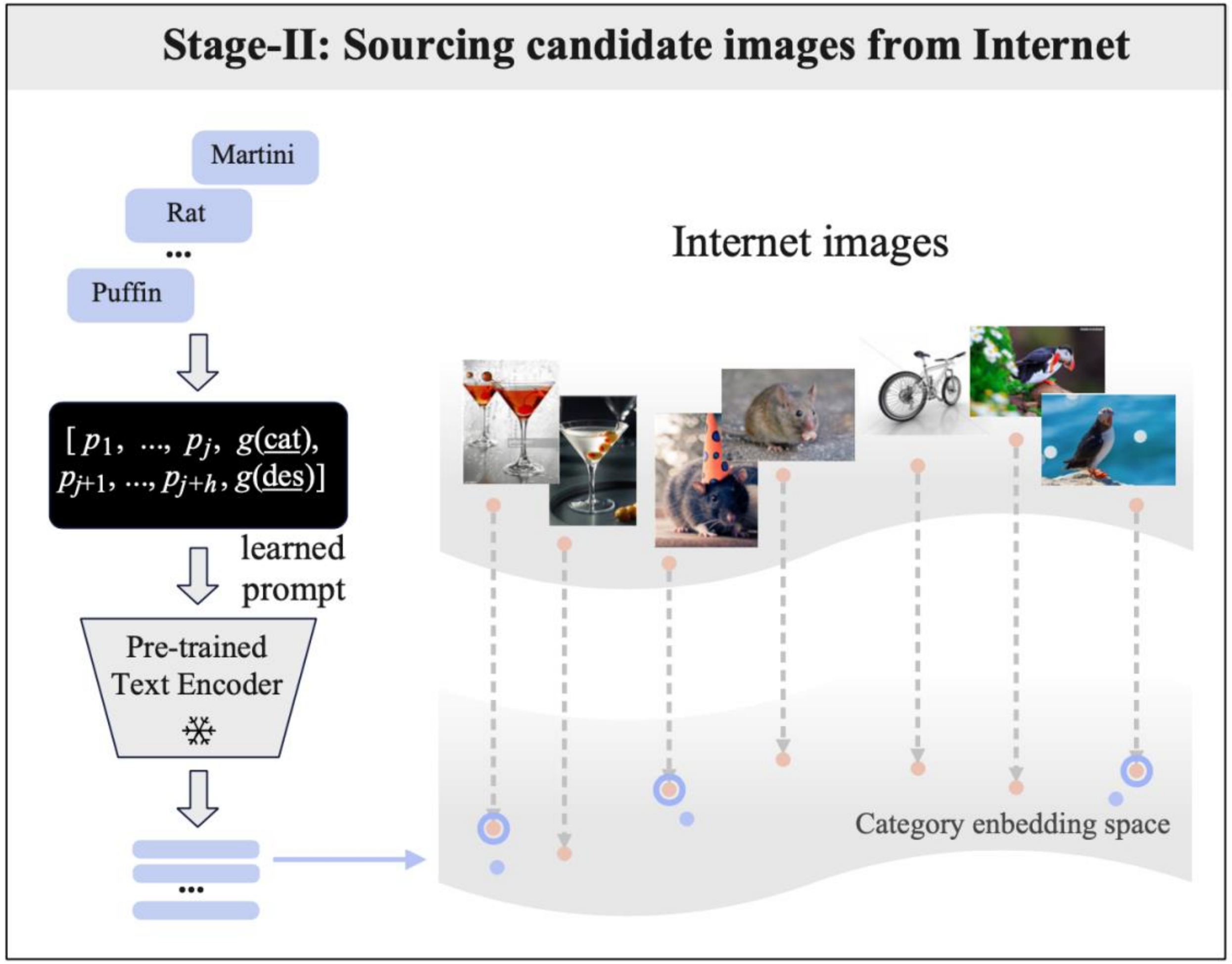
Table 2: Comparison on manually designed and learned prompt. Here, we only use two learnable prompt vectors in PRL, *i.e.* $[1+1]$ refers to using one vector for prefix, and one vector for suffix.

	Prompt	AP_{novel}	AP_c	AP_f	AP
"a photo of [category]"	manual	7.4	17.2	26.1	19.0
"a photo of [category], which is [description]"	manual	9.0	18.6	26.5	20.1
regional prompt learning	[1+1]	11.1	18.8	26.6	20.3

Self-training

- Sourcing candidate images
  - LAION-400M, initial corpus
  - N novel category prompt
  - Keep images with highest similarity
  - w/o GT box

- Alternate RPL and sourcing



Self-training

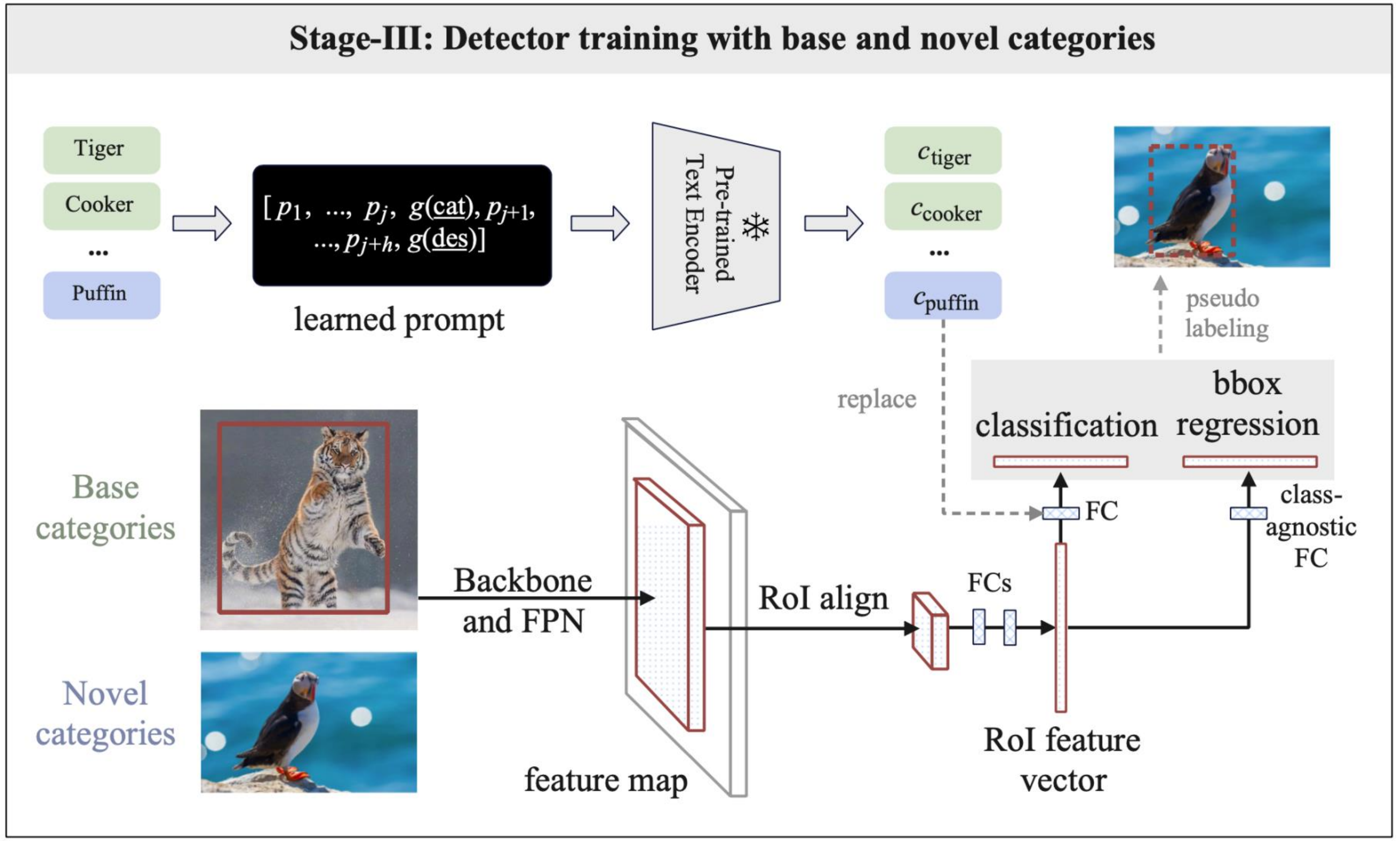
- Alternate RPL and sourcing

Table 3: Effectiveness of self-training with different prompts. 1-iter, 2-iter and 3-iter denote that Stage-I (*i.e.* RPL) and Stage-II (*i.e.* image sourcing) are performed for one, two or three iterations, respectively.

Prompt method	Self-training	AP_{novel}	AP_c	AP_f	AP
"a photo of [category], which is [description]"		9.0	18.6	26.5	20.1
	✓	15.3	17.7	25.8	20.4
Regional prompt learning		11.1	18.8	26.6	20.3
PromptDet (1-iter)	✓	15.9	17.6	25.5	20.4
PromptDet (2-iter)	✓	19.0	18.5	25.8	21.4
PromptDet (3-iter)	✓	19.3	18.3	25.8	21.4

Self-training

- Bounding box generation
  - Sourced images, object-centric
  - Top-K proposals, objectness
  - Maximal classification score
  - Re-training



Self-training

Fig. 4: Visualisation of the generated pseudo ground truth for the sourced images.

Self-training

Table 4: **Left**: the comparison on different box generation methods. **Right**: the effect on increasing the sourced candidate images.

Method	AP_{novel}	AP_c	AP_f	AP
w/o self-training	10.4	19.5	26.6	20.6
image	9.9	18.8	26.0	20.1
max-size	9.5	18.8	26.1	20.1
max-obj.-score	11.3	18.7	26.0	20.3
max-pred.-score (ours)	19.0	18.5	25.8	21.4

#Web images	AP_{novel}	AP_c	AP_f	AP
0	10.4	19.5	26.6	20.6
50	14.6	19.3	26.2	21.2
100	15.8	19.3	26.2	21.4
200	17.4	19.1	26.0	21.5
300	19.0	18.5	25.8	21.4

Dataset

Table 1: A summary of dataset statistics. The numbers in bracket refer to the number of base and novel categories.

Dataset	Train	Eval.	Definition	#Images	#Categories
LVIS	–	–	original LVIS dataset	0.1M	1203
LAION-400M	–	–	image-text pairs filtered by CLIP	400M	unlabeled
LVIS-base	✓	✗	common and frequent categories	0.1M	866
LAION-novel	✓	✗	image subset of novel categories	0.1M	337 (noisy)
LVIS *minival*	✗	✓	standard LVIS validation set	20K	1203 (866+337)

Result

Table 6: Detection results on the LVIS v1.0 validation set. Both Detic and our proposed approach have exploited the external images. However, in Detic, the images are manually annotated and thus indicated by '*'. Notably, PromptDet does not require a knowledge distillation from the CLIP visual encoder at the detector training, which is shown to prominently boost the performance but significantly increase the training costs.

Method	Epochs	Scale Jitter	Input Size	#External	AP_{novel}	AP_c	AP_f	AP
ViLD-text [11]	384	100~2048	1024×1024	0	10.1	23.9	32.5	24.9
ViLD [11]	384	100~2048	1024×1024	0	16.1	20.0	28.3	22.5
ViLD-ens. [11]	384	100~2048	1024×1024	0	16.6	24.6	30.3	25.5
Detic [36]	384	100~2048	1024×1024	1.2M*	17.8	26.3	31.6	26.8
PromptDet	12	640~800	800×800	0.1M	19.0	18.5	25.8	21.4
PromptDet	72	100~1280	800×800	0.1M	**21.4**	23.3	29.3	25.3