Aligning Bag of Regions for Open-Vocabulary Object Detection Supplemental Material

Size Wu[1] Wenwei Zhang[1] Sheng Jin[2,3] Wentao Liu[3,4] Chen Change Loy[1*]
[1]S-Lab, Nanyang Technological University [2] The University of Hong Kong
[3] SenseTime Research and Tetras.AI [4] Shanghai AI Laboratory
{size001, wenwei001, ccloy}@ntu.edu.sg {jinsheng, liuwentao}@sensetime.com

Table S1. Number of linear layers (#Layers) mapping region features to pseudo-words

#Layers	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
1	34.0	60.4	53.5
2	33.9	60.5	53.5
3	34.1	60.8	53.8

S1. Implementation Details

We provide more details of the implementation of BARON on OV-COCO [3] and OV-LVIS [2] benchmarks.

Sampling. For neighborhood sampling strategy, we obtain top K region proposals from the RPN and filter out those with an objectness score lower than 0.85. We also discard regions with an aspect ratio smaller than 0.25 or larger than 4.0. And regions with an area ratio smaller than 0.01 are also discarded. Then we apply NMS on the region proposals with IOU threshold 0.1. The region proposals after NMS are used for neighborhood sampling. We sample G bags of regions for each region proposal with a probability 0.3 to sample each surrounding candidate box. For OV-COCO, we set $K = 300$ and $G = 3$. For OV-LVIS, we set $K = 500$ and $G = 4$ due to the denser spatial distribution of object boxes in the LVIS dataset.

Classification Loss. We use CE loss as the classification loss $\mathcal{L}_{\text{cls}}$ on base categories. Given C object categories, we obtain the embedding f_i for the name of the i-th category by the text encoder ($\mathcal{T}$) of the VLM. We also learn a background embedding for non-object regions. If a region is labeled as the c-th category, the classification loss is

$$\mathcal{L}_{\text{cls}} = -\log \frac{\exp(\tau_{\text{cls}} \cdot \langle \mathcal{T}(w), f_c \rangle)}{\sum_{i=0}^{C} \exp(\tau_{\text{cls}} \cdot \langle \mathcal{T}(w), f_i \rangle)}, \tag{1}$$

where τ_{cls} is the temperature to re-scale the cosine similarity, f_C is the background embedding and w is the embedding (pseudo words) of the region. On OV-COCO, we set $\tau_{\text{cls}} = 50.0$. And on OV-LVIS, we set $\tau_{\text{cls}} = 100.0$ since there are orders of magnitude more categories defined in the LVIS dataset.

*Corresponding author.

Table S2. Different sampling strategies

#	Strategy	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}	#regions
1	Grid	25.4	58.0	49.5	36
2	Random	27.3	53.3	46.5	36
3	Random-Tight	29.5	56.9	49.7	36
4	Random-Neighbor	30.7	56.9	50.0	36
5	Ours (reduced)	**32.2**	58.3	51.5	36
6	Ours	**34.0**	60.4	53.5	216

Alignment Loss. Assuming there are G bags of regions and the image (teacher) and text (student) embeddings for the k-th bag of regions are f_v^k and f_t^k, the alignment loss $\mathcal{L}_{\text{bag}}$ on bag of regions is calculated as

$$\mathcal{L}_{\text{bag}} = -\frac{1}{2} \sum_{k=0}^{G-1} (\log(p_{t,v}^k) + \log(p_{v,t}^k)). \tag{2}$$

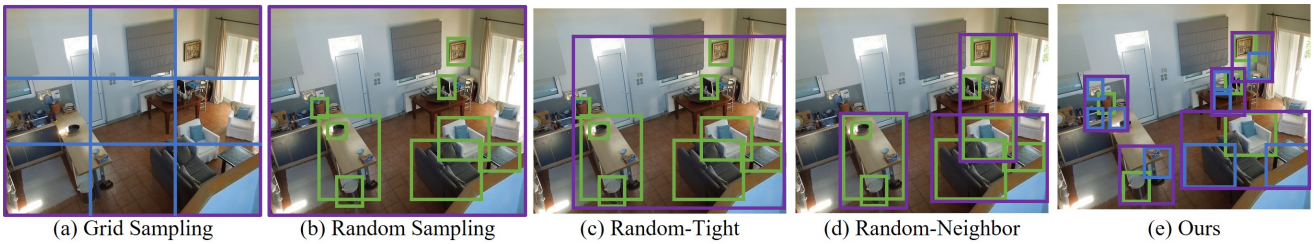
The $p_{t,v}^k$ and $p_{v,t}^k$ are calculated as

$$p_{t,v}^k = \frac{\exp(\tau_{\text{bag}} \cdot \langle f_t^k, f_v^k \rangle)}{\sum_{l=0}^{G-1} \exp(\tau_{\text{bag}} \cdot \langle f_t^k, f_v^l \rangle)} \tag{3}$$

$$p_{v,t}^k = \frac{\exp(\tau_{\text{bag}} \cdot \langle f_v^k, f_t^k \rangle)}{\sum_{l=0}^{G-1} \exp(\tau_{\text{bag}} \cdot \langle f_v^k, f_t^l \rangle)}, \tag{4}$$

respectively, where τ_{bag} is the temperature to re-scale the cosine similarity. Assuming there are totally N regions and the image (teacher) and text (student) embeddings for the k-th region are g_v^k and g_t^k, the alignment loss $\mathcal{L}_{\text{individual}}$ on individual regions is calculated as

$$\mathcal{L}_{\text{individual}} = -\frac{1}{2} \sum_{k=0}^{N-1} (\log(q_{t,v}^k) + \log(q_{v,t}^k)). \tag{5}$$



(a) Grid Sampling (b) Random Sampling (c) Random-Tight (d) Random-Neighbor (e) Ours

Figure S1. Comparison of different sampling strategies. Green boxes denote the region proposals. Blue boxes stand for sampled region boxes. The purple box represents the image crop of a bag of regions (a region group).

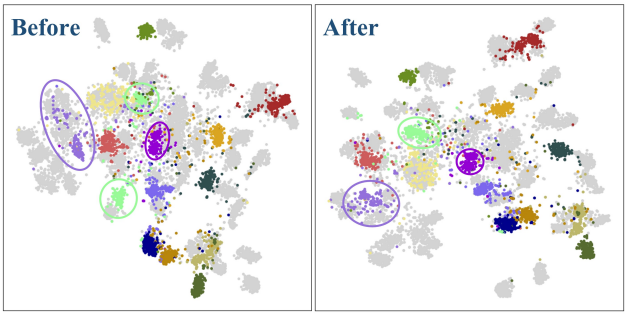


Figure S2. tSNE visualization of embeddings on COCO categories. **Left:** the region features *before* being projected to pseudo words. **Right:** embeddings *after* sending pseudo words to the text encoder.

The $q^k_{t,v}$ and $q^k_{v,t}$ are calculated as

$$q^k_{t,v} = \frac{\exp(\tau_{\text{individual}} \cdot \langle g^k_t, g^k_v \rangle)}{\sum_{l=0}^{N-1} \exp(\tau_{\text{individual}} \cdot \langle g^k_t, g^l_v \rangle)} \tag{6}$$

$$q^k_{v,t} = \frac{\exp(\tau_{\text{individual}} \cdot \langle g^k_v, g^k_t \rangle)}{\sum_{l=0}^{N-1} \exp(\tau_{\text{individual}} \cdot \langle g^k_v, g^l_t \rangle)}, \tag{7}$$

respectively, where $\tau_{\text{individual}}$ is the temperature to re-scale the cosine similarity.

On OV-COCO, we set $\tau_{\text{bag}} = 30.0$ and $\tau_{\text{individual}} = 50.0$. Since there are finer-grained definition of categories and denser distribution of object boxes in the LVIS dataset, we set $\tau_{\text{bag}} = 20.0$ and $\tau_{\text{individual}} = 30.0$ on OV-LVIS to make the contrastive learning harder.

Mapping Region Features to Pseudo-words. In our implementation, we used a single linear layer to map region features from the detector to pseudo-words. In Table S1, we show that adding more linear layers (#Layers) brings no noticeable improvements. This observation is also in line with Maaz *et al*. [5] that visual properties can be transferred to language models (LMs) by linearly mapping visual features to the input space of LMs.

Random Word Dropout. As we apply two different supervision to the pseudo words, the training can lead certain words to overfit to certain losses. To alleviate overfitting, we borrow the idea of Dropout [8] in neural networks where neurons are randomly dropped during training to avoid overfitting to specific neurons. We randomly discard pseudo words for each region with a probability p_{drop}. By default, we set $p_{\text{drop}} = 0.5$ for training on both OV-COCO and OV-LVIS.

Suppression on Novel Categories. On OV-COCO, we observe a tendency to overfit on base categories due to the smaller number of categories. And compared with OV-LVIS where the tail categories act as the novel categories, the distribution of novel and base categories on COCO is more balanced. We adopt the following strategies to alleviate suppression on novel categories: (1) detach the objectness prediction branch so that the suppression onto novel categories would not be back-propagated to the backbone; (2) save the sampled region proposals into a cache so that regions covering potential novel categories detected in certain iteration can be preserved throughout the training phase; (3) use the output of the second last layer of the VLM (CLIP) for classification and the final output for aligning bag of regions to reduce the competition between the two types of losses.

Figure S3. Image-guided inference of the detector trained on LVIS dataset. BARON can even recognize the cartoon characters in the reference images ('pikachu' and 'winnie pooh').

S2. Sampling Strategy

We have introduced two baseline sampling strategies, *i.e*. grid sampling and random sampling. The grid sampling strategy is to equally split an image into grids like the pre-training stage in OVR-CNN [9]. And the random sampling

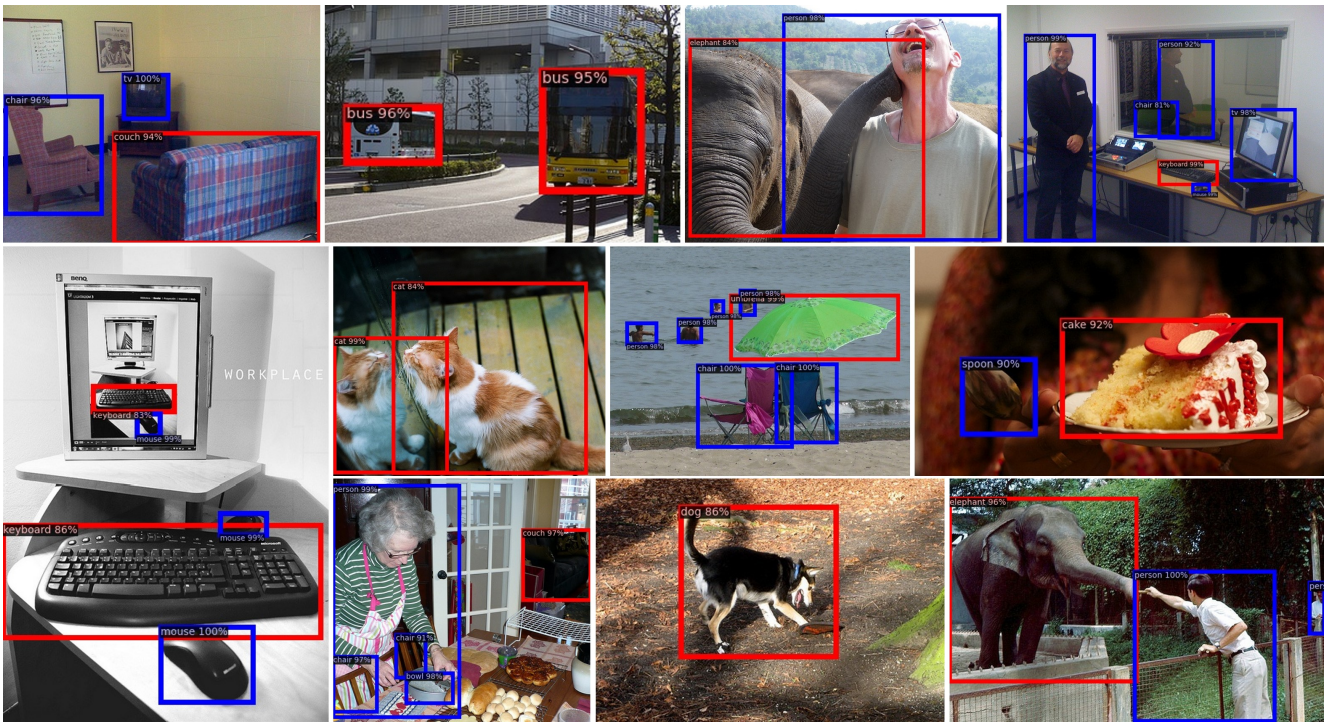


Figure S4. Visualization of detection results on OV-COCO. Red boxes are for novel categories, while blue boxes are for base categories.

strategy is to randomly sample region proposals to form a bag of regions. These two baseline strategies let the bag of regions represent the whole image. We add two other strategies to shift the focus to neighboring (local) regions.

We start from the random sampling strategy and let the bag of regions represent the image crop that tightly encloses them instead of the whole image (dubbed as Random-Tight). Then we move to the neighborhood centered on region proposals (dubbed as Random-Neighbor). For each center region proposal, we randomly sample 2 nearby region proposals with GIOU larger than 0.5 to make a bag of regions. We randomly take 12 region proposals as centers so that the total number of regions is 36, ensuring a fair comparison with other strategies. Table S2 shows the performance of these strategies.

In Fig S1, we show how these sampling strategies differ and how it gradually develops to our final option. In (a), we find the equally split grids may either contain too many objects or only small parts of an object. From (b) to (c), the bag of regions gradually shift to representing neighboring local regions from representing the whole image. However, we observe that there is always box size imbalance such as the left bottom bag of regions in (d). And there are also large area of redundant image contents between the regions in a bag as shown in (c). The box size imbalance and the redundant image contents hinder the image encoder of a VLM to effectively represent a bag of regions. As shown in (e), our sampling strategy obtains a bag of neighboring regions of equal size while capturing potential objects. Although we still observe image contents between sampled regions that do not belong to a bag of regions, they only account for a small portion of the image crop enclosing the bag of regions.

S3. Pseudo Word Encoding

Projecting visual features to word embedding space is common in region-based visual-language representation learning methods [1, 4]. In BARON, we project region features into pseudo words to fully exploit the inherent compositional structure of multiple semantic concepts and obtain more distinctive feature embeddings. In Fig S2, we show the tSNE visualization of the region features *before* being projected to pseudo words and embeddings *after* sending pseudo words to the text encoder (TE), *i.e.* $\mathcal{T}(w)$. Gray points represent base categories while chromatic points represent novel categories. With pseudo words encoded by TE, the categories are split into clusters of a more diverse distribution and distinct boundaries.

S4. Image-Guided Inference

We further examine the generalization ability of our method by using images to guide the inference of the detector. We use the image encoder of CLIP [6] to encode the reference image. And the detector used in this experiment

Figure S5. Visualization of detection results on OV-LVIS dataset. Red boxes and masks are for novel (rare) categories, while blue boxes and masks are for base categories.

Figure S6. Visualization of transfer detection results on Objects365 dataset.

is trained on the LVIS dataset. Given a reference image, our detector is able to detect the object in the reference image as shown in Fig S3. Our detector can even recognize the cartoon characters in the reference images ('pikachu' and 'winnie pooh').

S5. Detection Results

We show more detection results of our method in Fig S4 and Fig S5. On COCO dataset, BARON correctly detects novel categories including bus, keyboard, couch and so on.

On LVIS dataset, BARON detects rare categories like salad plate, fedora hat, gas mask and so on. We also visualize the results when transferring the LVIS-trained detector to Objects365 [7] dataset in Fig S6. We find that the LVIS-trained detector is able to correctly recognize a wide range of object concepts defined in Objects365 dataset, exhibiting impressive generalization ability.

S6. Potential Negative Societal Impacts

Our models have learned knowledge from vision-language models (VLMs) that are pre-trained on large-scale web image-text pairs. They potentially inherit and even reinforce harmful biases and stereotypes in the pre-trained VLMs. We suggest scrupulous probing before applying our models for any purpose.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Eur. Conf. Comput. Vis.*, pages 104–120, 2020. 3

[2] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 1

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Adv. Neural Inform. Process. Syst.*, 2019. 3

[5] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *Eur. Conf. Comput. Vis.*, pages 512–531, 2022. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021. 3

[7] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Int. Conf. Comput. Vis.*, 2019. 5

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014. 2

[9] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2