Efficient Feature Distillation for Zero-shot Annotation Object Detection: Supplementary Material

Zhuoming Liu[*], Xuefeng Hu[*], Ram Nevatia[†]
University of Southern California
{liuzhuom, xuefengh, nevatia}@usc.edu

1. Implementation Details

We include all implementation details in this section.

1.1. Adapt Image-language Model Feature

We use the publicly available pretrained CLIP [5] model ViT-B/32 as the open-vocabulary classification model, with an input size of 224×224.

Based on the detection setting we use for training and evaluating our detector, we adapt the CLIP to two detection domains: COCO [4] detection domain, LVIS detection domain [2]. We finetune the layer normalization layers in the CLIP with base category instances in COCO or LVIS based on the detection setting we use and maintain all other parameters fixed. All base category instances are cropped by 1.2x enlarged GT bboxes. We conduct the zero padding to convert each cropped region to the square and apply the default preprocessing pipeline of the CLIP.

We use CLIP to predict the category of each cropped region and calculate the cross-entropy loss with the GT label of each region. We finetune the model by optimizing the Cross-Entropy Loss. We use AdamW optimizer with a learning rate of 0.0001, batch size 4 and clip the L2 norm of the gradients when larger than 0.1. We finetune the model for 12 epochs.

1.2. Generate CLIP Proposals

When generating the CLIP Proposals, we still use the CLIP model we mentioned in section 1.1 as a classifier to select the distillation regions. If we will use the adapted CLIP's feature to train the detector, we will use the adapted CLIP to generate the CLIP Proposals. Otherwise, we use the unadapted CLIP to generate CLIP Proposals.

We generate the CLIP proposals on all the training images of the detection dataset base on the detection setting we use. We first resize the image with the image ratio maintained. The long edge of the image will be resized into 1333 as width or 800 as height.

We generate the anchors on each image with a stride of 32 pixels and with 5 different sizes (32, 64, 128, 256, 512), and 3 different ratios (1:1, 2:1, 1:2). We select the top 1000 anchors after NMS as CLIP Proposals on each image. We filter out the anchors which have high IoU with the base category GT bboxes to reduce the redundancy since we will add 1.2x enlarged base category GT bbox as part of the CLIP Proposals. In model training, we randomly select a fixed subset with 200 CLIP Proposals on each image for training.

1.3. Detection Setting

In COCO detection setting, the dataset is divided into 48 base categories and 17 novel categories. 15 categories without a synset in the WordNet hierarchy are removed.

We filter out the training images which do not have base category annotation. Following the setting in [8], we filter out the images that have neither the base category instances nor the novel category instances in the validation set. The training set contains 107761 images and 665387 base category instances. The validation set contains 4836 images and 28538 base category instances and 33152 novel category instances. We evaluate the model in a generalized setting, which evaluates the base and novel categories at the same time. AP50 is used as the evaluation metric.

In LVIS detection setting, the dataset is divided into 866 base categories (containing 405 frequent categories and 461 common categories) and 337 novel categories (337 rare categories). Our LVIS-Fbase split uses the frequent categories as the base(405 categories), common and rare categories as the novel(common has 461 categories, rare has 405 categories). The training set contains 98531 images and 1200258 base category instances. The validation set contains 19442 images and 230427 base category instances and 14280 novel category instances. We aggregate the model performance in frequent, common, and rare categories separately. AP is used as the evaluation metric.

[*] Equal contribution.
[†] Corresponding author

2. Experiments in Few-shot Detection Settings

In few-shot object detection, the model is trained on the base category's annotations and evaluated on novel categories. The only difference is that in few-shot detection, each novel category has the same number of annotated objects(i.e, K-shot), which can be used to improve the model performance on the novel before the model is evaluated. We directly evaluate our model in the few-shot benchmark, without using this K-shot additional information.

Datasets and Evaluation Metrics. We evaluate our approach on PASCAL VOC 2007+2012 and COCO. For the few-shot PASCAL VOC dataset, we combine the trainval set of 2007 with the one of 2012 as training data. PASCAL VOC 2007 test set is used for evaluation. The 20 classes are divided into 15 base classes and 5 novel classes. We evaluate our model in three different base/novel splits used in [7]. Split 1 has 14631 training images with 41084 base category instances, and the validation set has 4952 images, 10552 base category instances, and 1480 novel instances. Split 2 has 14779 training images with 40397 base category instances, and the validation set has 4952 images, 10447 base category instances, and 1585 novel instances. Split 3 has 14318 training images with 40511 base category instances, and the validation set has 4952 images, 10605 base category instances, and 1427 novel instances.

For the few-shot COCO dataset, we use the COCO train2017 as training data and evaluate our model on the COCO val2017. The 20 categories that exist in PASCAL VOC are used as the novel categories, while the rest of the 60 categories are used as the base categories. The training set has 98459 images and 367189 base category instances. The validation set has 5000 images and 15831 base category instances and 36781 novel category instances.

AP50 is used as the evaluation metric in PASCAL VOC, while AP and AP50 are used in COCO.

Model. Following previous work in few-shot detection, we train a Faster R-CNN [6] model with ResNet-101 FPN backbone. The backbone is pretrained on ImageNet. We use SGD as the optimizer with batch size 4, learning rate 0.005, momentum 0.9, and weight decay 0.0001. We also adopt linear warmup for the first 500 iterations, with a warm up ratio is 0.001. We apply multi-scale train-time augmentation. For the PASCAL VOC dataset, we train the model for 21 epochs and divide the learning rate by 10 at epoch 15 and epoch 18. For the COCO dataset, we train the model for 18 epochs and divide the learning rate by 10 at epoch 14 and epoch 16.

Baselines. We compare EZAD's performance with two few-shot detection models, TFA [7] and Meta Faster R-CNN [3] as the baselines. The TFA model with linear layer as the classifier is noted as *TFA w/fc*, while the model with cosine classifier is noted as *TFA w/cos*.

Results. Table 1 shows the results on the PASCAL dataset. EZAD achieves 40.9% in novel AP50 averaged over three different splits. EZAD's performance matches the TFA 3-shot performance in split1 and split2 and is 4.7% higher than TFA in split3. Compared with the TFA's performance on base, EZAD is 1.8% higher. For Meta Faster R-CNN, it generates proposals for each category on each image, which needs multiple forward passes. Its inference time will be much slower if the dataset has a large number of novel categories. Compared with the Meta Faster R-CNN, EZAD outperforms it without using any additional annotations by a 1.6%, 3%, and 6.9% in three different splits, respectively. Table 2 shows the results on the COCO dataset. EZAD achieves 10.2% and 22.2% in AP and AP50, respectively, matching TFA's 10-shot performance and 2.6% and 5.9% higher than the Meta Faster R-CNN's 2-shot performance in AP and AP50, respectively. Our model zero-shot performance on the few-shot setting shows the power of adapted multi-modal feature space and validates the effectiveness of using CLIP Proposals as distillation regions.

Method	Shot	Novel AP50			
		Split1	Split2	Split 3	Avg
TFA w/fc	1	36.8	18.2	27.7	27.6
TFA w/fc	2	29.1	29.0	33.6	30.6
TFA w/fc	3	43.6	33.4	42.5	39.8
TFA w/cos	1	39.8	23.5	30.8	31.4
TFA w/cos	2	36.1	26.9	34.8	32.6
TFA w/cos	3	**44.7**	**34.1**	42.8	40.5
MF R-CNN	1	43.0	27.7	40.6	37.1
Ours	0	44.6	30.7	**47.5**	**40.9**
Split1 Base(AP50): TFA (3-Shot)=79.1, Ours=80.8					

Table 1. Evaluation results on the novel categories of PASCAL VOC few-shot benchmark. MF R-CNN means Meta Faster R-CNN. Our model zero-shot performance on the novel match the TFA's performance in its 3-shot setting. Our model also has a better performance on base.

Method	Shot	AP	AP50
TFA w/fc	10	10.0	19.2
TFA w/cos	10	10.0	19.1
MF R-CNN	2	7.6	16.3
Ours	0	**11.0**	**23.5**

Table 2. Evaluation results on novel categories of COCO few-shot benchmark. MF R-CNN means Meta Faster R-CNN. Our model zero-shot performance on the novel match the TFA's performance in its 10-shot setting.

3. Additional Ablation Study

Table 3 presents the experimental results on how the size of the bounding box (bbox) that we use to crop the in-

Bbox Size	General			
	L	M	S	Avg
0.8x GT	62.3	54.0	23.2	47.1
1.0x GT	**64.0**	61.9	32.9	53.4
1.2x GT	61.3	**62.2**	36.9	**53.9**
1.5x GT	56.7	59.5	40.6	52.6
2.0x GT	50.5	52.6	**42.9**	48.9

Table 3. The classification accuracy (ACC) of the unadapted CLIP on COCO instances with different sizes of the GT bboxes to crop the instances. We decide to use the 1.2x enlarged GT bbox to crop the instance since it has the best average ACC.

Epoch	Distill Region	Base	Novel	Overall
12	RPN Proposal	56.9	24.6	48.5
12	CLIP Proposal	55.7	30.4	49.0
36	RPN Proposal	**60.2**	24.3	50.8
36	CLIP Proposal	59.9	**31.6**	**52.1**

Table 4. Ablation study on using CLIP Proposals as distillation in COCO benchmark. The model trained with CLIP Proposals has much better performance on novel categories.

stances in the COCO [4] dataset affects the classification accuracy (ACC) of the unadapted CLIP [5]. For the large objects, the more accurate bbox provided the higher ACC CLIP can achieve. For the small objects, CLIP needs more background information to be correctly classified. In all settings, the average ACC over all three sizes of the bbox is still much lower than the classifier of the well-trained detector, indicating the domain gap between the training data of the CLIP and the detection dataset exists. We use the 1.2x GT bbox to crop the base GT instance since it has the highest average ACC.

We provide an additional ablation study in Table 4. We train all models with the adapted CLIP features. For the models trained with 12 epochs, the performance on novel categories of the model trained with the RPN proposals is 5.8% lower than the one of the model trained with the CLIP proposals, though the former has slightly better performance on base categories. For the models trained with 36 epochs, two models (RPN proposal and CLIP proposal) has similar performance on base categories, and the model trained with the CLIP proposal features still have much better novel category performance. This indicates that the negative effect on model performance on base categories caused by the CLIP proposal is negligible and can be alleviated by a longer training schedule. It also shows that the information of base categories provided by the distillation has redundancy, which may accelerate the model convergence on base, but may not improve the model performance.

4. Additional Visualizations

Fig 1 shows the visualization of using CLIP Proposals and RPN proposals as distillation regions in the COCO setting. The blue boxes and green boxes represent the GT bboxes of the novel and base categories. The red boxes represent the CLIP Proposals or the RPN proposals with the highest IoU with the novel GT bboxes. The three images on the left show that the CLIP Proposals can cover most of the novel category objects although the boxes may not accurate, while the RPN regards some of the novel objects as background and just ignores them. Although the CLIP proposals are not accurate, the features extracted from these boxes are accurate and meaningful. This phenomenon is also proved by the experiments in [1]. Therefore, using the CLIP Proposals as distillation regions provides more novel category information and improve the detector's performance on the novel.

Fig 2 shows the tSNE embeddings of the COCO instance features of the unadapted CLIP and the adapted CLIP. We collect 20 GT instances for each base and novel category in COCO setting and extract their features from unadapted CLIP or adapted CLIP, and then generate the tSNE embeddings with these features. The GT instances in the adapted CLIP feature space form some dense clusters. This indicates that the CLIP's feature space has been adapted in the COCO dataset domain and the features become more discriminating after adaptation, improving the classification accuracy. The dots do not form a dense cluster mostly come from the "person" category. Since the instances of the person usually show up with other categories instances and occluded by other objects, therefore the person categories features are more scattered.

References

[1] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[3] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022. 2

[4] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

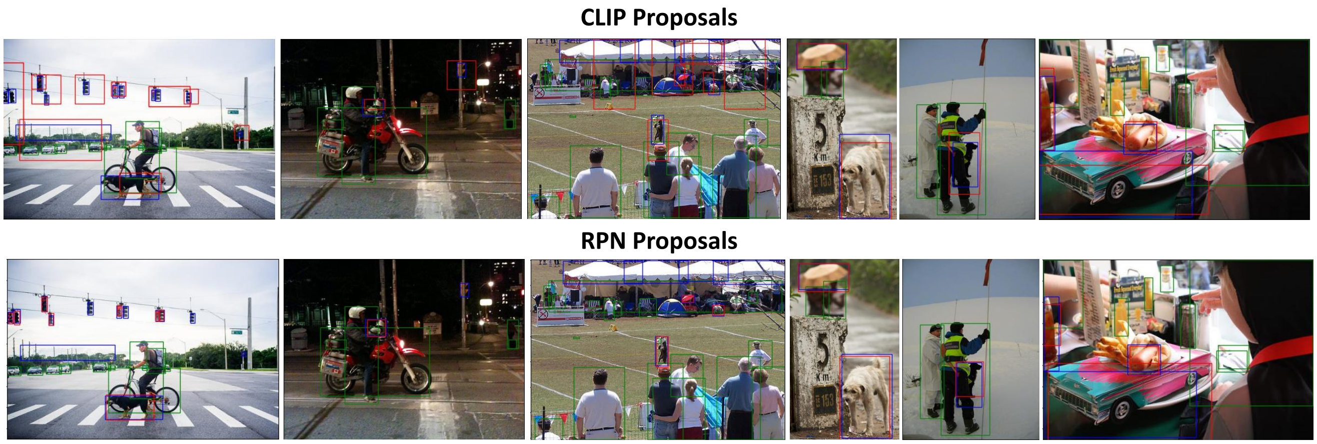

Figure 1. Visualization of using CLIP Proposals or RPN proposals as distillation regions in COCO setting. The blue boxes and green boxes represent the GT bboxes of the novel and base categories. The red boxes represent the CLIP proposals or the RPN proposals with the highest IoU with the novel GT bboxes. The visualization shows the CLIP proposals can cover more novel objects even though the box may not accurate.

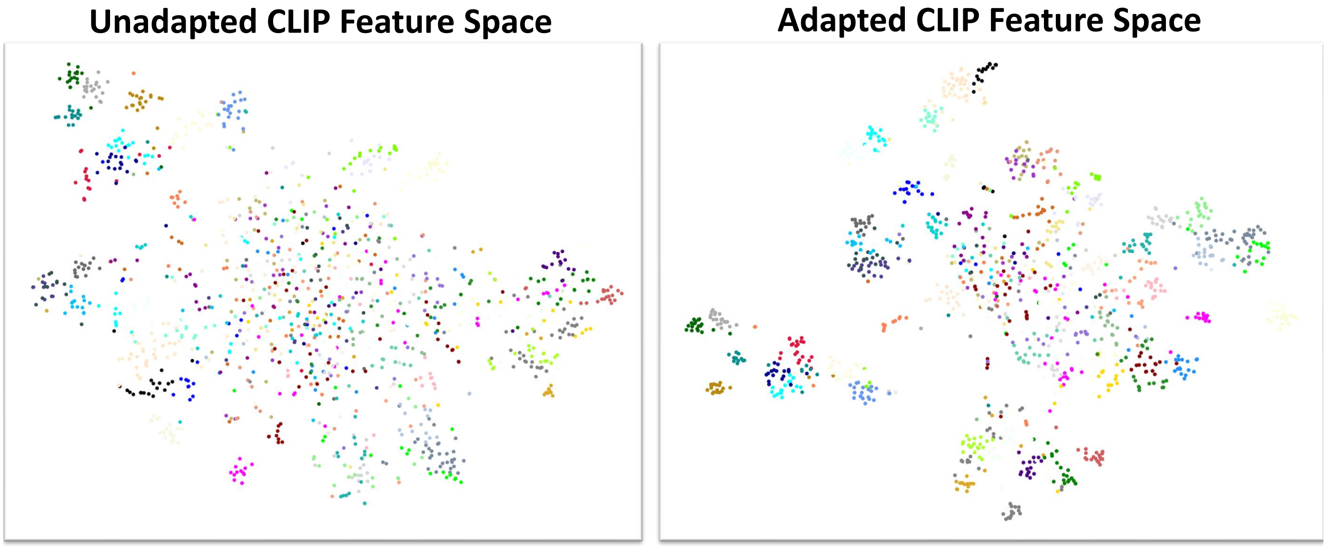

Figure 2. The tSNE embeddings of the COCO GT instance feature from the unadapted CLIP and adapted CLIP. The GT features from the adapted CLIP form more dense clusters, indicating that the features become more discriminating and the CLIP is adapted into the detection dataset domain.

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[7] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 2

[8] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1