SUPPLEMENTARY: REVISITING FEW-SHOT OBJECT DETECTION WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

1 ANALYSING PROPOSED SETUP FOR K=5,30 SHOTS

For K=5 shots, we note that naive fine-tuning barely improves the overall AP (See Table 1). Incorporating negatives in this data-scarce setting helps us reach close to the true negatives upper bound (15.61 vs 15.85 AP). We can see that nuImages is a challenging dataset as even with all annotations, we are able to improve only by ~ 3.8 points over the zero-shot numbers. We see a decrease in performance for the `Few` group with all annotations. This is because we re-introduce the long-tailed behavior which artificially did not exist in FSOD.

As we have a lot more positive annotations in K=30 shot setting, we see diminishing returns of leveraging negatives to improve performance (See Table 2). Naive fine-tuning improves by almost ~ 3 AP points over the zero-shot performance (17.20 vs 14.26 AP). Pseudo-negatives still help improve the overall AP, but the difference is ~ 0.1 AP. The exhaustive annotations show a big jump in `Many` group performance. For example, we have over 1300 annotations of car in this setting vs 30 for FSOD, which explains the boost in scores for `Many`.

Approach	**5 shots**: Average Precision (AP)			
	All	Many	Medium	Few
Detic (Zhou et al., 2022) Zero-Shot	14.26	27.28	16.88	2.36
+ fine-tuning	14.64	25.70	18.26	**3.22**
+ FedLoss	14.70	26.88	18.23	2.45
+ Inverse FedLoss	15.02	27.29	18.38	2.89
+ Pseudo-Negatives	**15.61**	**28.88**	**18.59**	3.12
w/ True Negatives	15.85	29.45	18.62	3.28
w/ Exhaustive Annotations	17.05	32.43	20.13	2.83

Table 1: **Analysis of 5-shot performance on nuImages**. We observe similar trends to 10-shots as we see improving numbers when we handle negatives using InvFedLoss and Pseudo-Negatives. More specifically, we get very close to the upper bound with True Negatives which shows how pseudo-negatives can really boost performance in a very limited data setting.

Approach	**30 shots**: Average Precision (AP)			
	All	Many	Medium	Few
Detic (Zhou et al., 2022) Zero-Shot	14.26	27.28	16.88	2.36
+ fine-tuning	17.13	28.97	21.96	**4.10**
+ FedLoss	16.52	29.66	21.22	2.53
+ Inverse FedLoss	17.10	29.35	**22.02**	3.63
+ Pseudo-Negatives	**17.20**	**30.48**	21.60	3.37
w/ True Negatives	18.21	30.88	23.03	4.61
w/ Exhaustive Annotations	19.43	34.37	23.90	4.31

Table 2: **Analysis of 30-shot performance on nuImages** We see that when increasing the number of shots, the gains from our method reduce. This can be attributed to the fact that adding more positive information is much more meaningful to the detector, and therefore we see such improvements that are overshadowed by any gains we see from negatives.

2 IMPLEMENTATION AND TRAINING DETAILS

2.1 LVIS V0.5

We select Detic with a Resnet-50 backbone as our architecture of choice to have an even comparison with prior methods. We conduct pre-training on `LVIS-base` for $90k$ iterations with a batch size of 32, and a learning rate of 0.0002 is used with the AdamW optimizer (Loshchilov & Hutter, 2017). Image size of 640×640 is used and we also enable Repeat Factor Sampling (Gupta et al., 2019).

For fine-tuning, we sample upto 10 shots for each class in LVIS following (Wang et al., 2020). We use a batch size of 32, learning rate of 2.5×10^{-5} for $46k$ iterations. We do not use Repeat Factor Sampling for fine-tuning. For the Federated Loss and Inverse Federated Loss experiments, we sample 50 categories for each training image, i.e $|S| = 50$. For the pseudo-negatives experiments, we derive negatives from pseudolabels with atleast 20% confidence.

2.2 NUIMAGES

To showcase results on our proposed setup with nuImages, we adopt a pretrained model which is trained on more than one dataset. Specifically, we select Detic with a Swin-T backbone, pre-trained on `LVIS+COCO` and ImageNet-21k data. We use an image size of 1600×900, batch size of 8 and learning rate of 3.75×10^{-6} for the AdamW optimizer. We conduct this fine-tuning for 8000 iterations. For federated and inverse federated loss experiments we set number of sampled categories, i.e $|S| = 6$. This is a hyperparameter and can be further tuned for other datasets. Similar to LVIS, we use 20% confidence for deriving negatives from pseudolabels.

The original Detic implementation samples $|S|$ categories for a batch, rather than an image. This works for LVIS as there exist over 1200 categories such that you may get different categories at every sampling step. But we have 18 classes in nuImages, therefore we re-implemented the negative category sampling step to work for each image. This was needed to scale batch size lest we sample all categories for a batch every time.

3 LVIS V0.5 WITH SWIN BACKBONE

In this section we intend to evaluate the performance change that occurs when swap out the backbone. Specifically, we replace ResNet-50 backbone by a SWIN-B model in Detic. Note that this Detic model is also pre-trained on just `LVIS-base` and then finetuned on upto 10 shots of `LVIS-base` and `LVIS-rare`. On comparing just the zero shot numbers, we see a jump over 5 points in the overall AP. Regardless of changing the backbone, we see the same trend of improving numbers as we handle negatives. Using Pseudo-Negatives gives us our best numbers, which has a difference of ~ 13 points for AP_r compared to its ResNet counterpart. These results solidify our insights about the FSOD tasks and validate the techniques designed on top of it.

REFERENCES

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiawei Ma, Yulei Niu, Jincheng Xu, Shiyuan Huang, Guangxing Han, and Shih-Fu Chang. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3208–3218, 2023.

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.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022.

Approach	10-shots			
	AP	AP_f	AP_c	AP_r
ResNet-50 Backbone				
TFA w/ fc (Wang et al., 2020)	24.1	27.9	23.9	14.9
TFA w/ cos (Wang et al., 2020)	24.4	27.7	24.3	16.9
DiGeo (Ma et al., 2023)	24.9	28.5	24.6	17.3
Detic (`Base` Only) (Zhou et al., 2022)	30.0	34.4	30.8	16.3
+ Fine-Tuning (`Base` + `Novel`)	30.0	33.2	31.9	15.5
w/ FedLoss	30.8	33.9	32.7	17.4
w/ InvFedloss	31.1	34.3	32.5	18.7
w/ Pseudo-Negatives	**31.6**	**34.6**	**33.2**	**19.2**
Swin backbone				
Detic (`Base` Only) (Zhou et al., 2022)	35.2	38.7	36.8	21.4
+ Fine-Tuning (`Base` + `Novel`)	35.9	37.1	37.8	26.7
w/ FedLoss	36.5	36.7	38.3	30.4
w/ InvFedloss	37.1	**37.8**	**38.5**	31.1
w/ Pseudo-Negatives	**37.2**	37.7	38.2	**32.6**

Table 3: **LVIS FSOD with SWIN-Backbone** We evaluate the impact of changing backbones independent of large-scale pre-training. We adhere to the standard FSOD setup and pre-train Detic on `LVIS-base`. We observe a massive jump in performance across all categories! The best performing Swin model, outperforms the best ResNet model by 5.4 points in overall AP and ~ 13 points on AP_r. Regardless of the powerful SWIN backbone, we still see improvements by incorporating our insight about handling negatives in FSOD tasks.