A Closer Look at Cross-Domain Few-Shot Object Detection: Fine-Tuning Matters and Parallel Decoder Helps

Supplementary Material

Contents

A. Full training details

Hyperparameters. We list the hyperparameters of our training in Table 1. Specifically, patience without progressive fine-tuning in the plateau scheduler indicates that the patience used in the *Baseline (Naive aug & Plateau scheduler)* in the experiments in Table 4 (main paper), as well as the *baseline* and the *baseline + HED (w/ re-init Dn Query)* in Figure 2 (main paper) and Table 14. For the progressive fine-tuning, we set different patience for the first and second stages. The smaller patience in the first stage can prevent fast overfitting caused by the bigger learning rate at the beginning of the training, and the bigger patience in the second stage with a smaller learning rate can help the model to retain the good performance after the first plateau. When the model is trained without the progressive fine-tuning pipeline, the patience is set to the average patience for the whole training as a reasonable configuration. Note that we train 100 epochs for the sub-datasets, which is the same configuration as RF-DETR [10] used on RF100-VL full-shot benchmark [9], in order to make sure that the models are correctly and fully converged. This does not indicate that training all datasets requires this configuration, and one can set early stopping to save time. We apply the listed configurations for all the benchmark experiments listed in Tables 1-3 in the main paper without any dataset-specific tunings.

Computational resources. Our CD-FSOD [2] experiments are conducted with the MMGDINO-B model, and require 2 RTX 4090/3090 GPUs, or around 48GB GPU memory using other types of GPUs. For the ODinW-13 and RF100-VL experiments, the base model is MMGDINO-L, which is much bigger and requires 4 RTX 4090/3090 GPUs, or around 96GB GPU memory using other types of GPUs. The computational cost can be reduced to half if the batch size is set as 2 instead of 4 when the computational resources are limited.

B. Additional ablation study

The additional ablation experiments are based on CD-FSOD benchmark [2] with the pre-trained MMGDINO-B model.

B.1. Parallel decoder layers initialization

We conducted an ablation experiment on the initialization of the parallel decoder layer in HED to check if the pre-trained weights were still important for the parallel decoder layers.

The results are shown in Table 2. We observed that although the single-member in HED has fewer parameters than the conventional detection decoder and is fine-tuned on a small training set (which may make it more prone to overfitting), initializing this part with pre-trained weights remains crucial. Furthermore, if more parameters are randomly initialized (e.g., all parameters of the model), we found that the model becomes difficult to converge, i.e., transfer learning becomes impossible.

B.2. Learning rate scheduler

As proposed in the main paper, we argue that the plateau scheduler can better autonomously adjust the learning rate for fine-tuning on larger benchmarks (e.g., RF100-VL [9]), while multi-stage training can also be triggered according to the plateau. However, the cosine scheduler is more commonly used in object detection tasks and can also be applied with HED in our pipeline. Therefore, we conducted ablation experiments using a cosine scheduler. Specifically, we used a cosine scheduler instead of a plateau scheduler when using HED, employing both single-stage and progressive fine-tuning. When using the cosine scheduler, we enabled two-stage training at half of the training epochs.

The results are provided in Table 3. We observed that, when using the cosine scheduler, progressive fine-tuning is still more effective than single-stage. Secondly, the performance of the cosine scheduler is comparable to that of the plateau scheduler. Intuitively, we believe that since models are more prone to overfitting when fine-tuning with few samples, it is crucial to decrease the learning rate at the right time. Therefore, both schedulers are more suitable than the fixed conventional milestone scheduler. Furthermore, our goal is to find a more general pipeline, in which case the plateau scheduler is a better fit.

Component	Hyperparameters
Init. learning-rate for detection transformers	1e-4
Init. learning-rate for image and language backbones	2e-5
Minimum learning rate	1e-6
Weight decay	0.05
Batch size	4
Number of epochs	100
Plateau scheduler - Patience w/o progressive ft.	5
Plateau scheduler - Patience in the 1st stage w/ progressive ft.	3
Plateau scheduler - Patience in the 2nd stage w/ progressive ft.	8
Plateau scheduler - factor	0.5
Re-initialization ratio for denoising queries	0.5
Hybrid ensemble decoder structure	1-stacked layer + 5-parallel layers
Random flip probability	0.5
YOLOXHSVRandomAug probability	0.5
CatchedMixup probability	0.3

Table 1. **Hyperparameter list applied for training on all benchmark experiments.**

CD-FSOD | CD-Mixed	1-shot	5-shot	10-shot
Randomly Initialized	33.7 | 25.04	44.1 | 36.08	47.2 | 38.02
Pre-trained	**34.9 | 25.68**	**45.0 | 37.14**	**47.9 | 40.34**

Table 2. Fine-tuning results with and without randomly initialized parallel layers for HED.

CD-FSOD | CD-Mixed	1-shot	5-shot	10-shot
1-stage (cosine) + HED	34.6 | 24.92	43.6 | 33.52	47.2 | 38.30
2-stage (cosine) + HED	34.6 | 25.46	44.2 | 35.34	47.4 | 38.42
2-stage (plateau) + HED (Ours)	**34.9 | 25.68**	**45.0 | 37.14**	**47.9 | 40.34**

Table 3. Fine-tuning results for the model using cosine and plateau scheduler with and without our progressive fine-tuning strategies.

C. Full numerical results

C.1. Full results on ODinW-13

We present the full results on the ODinW-13 benchmark [11] in Tables 4 to 7, which correspond to the average results across three official seeds for the 1-shot, 3-shot, 5-shot, and 10-shot settings, respectively.

We observe that other open-source models achieve comparable zero-shot performance to our base model. However, by applying our proposed training strategies, our model consistently and significantly outperforms the other fine-tuned open-source models.

Furthermore, while GroundingDINO 1.5 Pro [8] is an updated version of GroundingDINO [7], featuring a much larger-scale pre-training dataset and a more powerful architecture, which obtained much better zero-shot performance than our base model, our methodology still succeeds in outperforming the fine-tuned version of this strong closed-source model. This result further demonstrates the effectiveness of our proposed method.

Datasets	1-shot
AerialMaritimeDrone	26.267
Aquarium	47.500
CottontailRabbits	76.967
EgoHands	67.133
NorthAmericaMushrooms	81.067
Packages	65.833
PascalVOC	68.400
Raccoon	70.600
ShellfishOpenImages	66.533
VehiclesOpenImages	69.667
pistols	73.267
pothole	29.233
thermalDogsAndPeople	78.467
Average	**63.1**

Table 4. **Average results of ODinW-13 of the three official seeds of runs** for 1-shot setting.

C.2. Full results on RF100-VL

The full fine-tuning results of our method on RF100-VL [9] are listed in Table 8 to 13. Based on the results of SAM3 [1] in the main paper, we argue that for few-shot object detection, powerful visual detection or segmentation foundation models have an advantage in the domains with more natural images, such as Flora-Fauna and even Aerial images. However, in specific domains, such as medical images and documentation content, the results of few-shot fine-tuning are not ideal. We believe this is partly due to the fact that the data distribution

Datasets	3-shot
AerialMaritimeDrone	34.033
Aquarium	52.767
CottontailRabbits	77.067
EgoHands	69.033
NorthAmericaMushrooms	83.567
Packages	67.733
PascalVOC	70.633
Raccoon	71.167
ShellfishOpenImages	68.133
VehiclesOpenImages	69.367
pistols	71.933
pothole	40.067
thermalDogsAndPeople	79.367
Average	**65.8**

Table 5. **Average results of ODinW-13 of the three official seeds of runs** for 3-shot setting.

Datasets	5-shot
AerialMaritimeDrone	38.200
Aquarium	53.733
CottontailRabbits	76.533
EgoHands	72.767
NorthAmericaMushrooms	87.000
Packages	70.833
PascalVOC	69.933
Raccoon	75.733
ShellfishOpenImages	67.733
VehiclesOpenImages	72.033
pistols	70.267
pothole	42.133
thermalDogsAndPeople	81.367
Average	**67.6**

Table 6. **Average results of ODinW-13 of the three official seeds of runs** for 5-shot setting.

Datasets	10-shot
AerialMaritimeDrone	39.367
Aquarium	55.167
CottontailRabbits	75.200
EgoHands	71.900
NorthAmericaMushrooms	88.600
Packages	73.033
PascalVOC	70.933
Raccoon	75.800
ShellfishOpenImages	67.700
VehiclesOpenImages	72.000
pistols	71.633
pothole	46.333
thermalDogsAndPeople	83.600
Average	**68.6**

Table 7. **Average results of ODinW-13 of the three official seeds of runs** for 10-shot setting.

Aerial Datasets	10-shot
aerial-airport	48.1
aerial-cows	34.4
aerial-sheep	48.2
apoce-aerial-photographs -for-object-detection- of-construction-equipment	46.2
electric-pylon-detection-in-rsi	21.1
floating-waste	36.5
human-detection-in-floods	36.2
sssod	53.6
uavdet-small	39.0
wildfire-smoke	49.9
zebrasatasturias	42.9
Average	**41.5**

Table 8. **Results on RF100-VL - Aerial category** for 10-shot Performance (mAP)

of these domains differs from the data during large-scale pre-training. For example, the training data for open-source MM-GroundingDINO [12] is mainly natural images from OpenImage [4], COCO [5], and so on. In this case, even if the zero-shot performance of these specific domains is at the same low level as the zero-shot performance with more natural images, after fine-tuning, the model still needs more data than natural images to achieve a higher detection accuracy in this specific domain.

C.3. Full results on CD-Mixed set

We provide numerical results on the proposed CD-Mixed set in Table 14, which is consistent with the results shown on Figure 3 in the main paper. The prediction average results are not consistent with the ones in the main paper, since we omitted the FISH dataset in the CD-Mixed set, as the FISH dataset has the same domain (underwater imagery) as the UODD dataset, and this might cause the overlapping object categories.

We first observe that introducing a large number of out-of-distribution (OOD) samples significantly reduces mAP, meaning that some OOD samples contain high-confidence detection results. However, since the samples in our chosen dataset have no overlap in terms of targets (e.g., underwater creatures will not appear in images of industrial defects, clip art will not appear in the images of underwater, and close-ups of insects will not appear in aerial photographs), these high-confidence samples are false positives, meaning results obtained by the model are due to overconfidence. This conclusion has also been found in COCO-FP [6] and aligns

Document Datasets	10-shot
activity-diagrams	30.4
all-elements	39.5
circuit-voltages	34.5
invoice-processing	23.7
label-printing-defect-version-2	54.1
macro-segmentation	34.1
paper-parts	38.8
signatures	68.9
speech-bubbles-detection	51.5
wine-labels	13.6
Average	**38.9**

Table 9. **Results on RF100-VL - Document category** for 10-shot Performance (mAP)

Flora-Fauna Datasets	10-shot
aquarium-combined	52.6
bees	30.9
deepfruits	64.9
exploratorium-daphnia	26.7
grapes-5	39.3
grass-weeds	46.7
gwhd2021	25.3
into-the-vale	64.4
jellyfish	31.8
marine-sharks	30.2
orgharvest	11.3
peixos-fish	30.1
penguin-finder-seg	73.4
pig-detection	47.0
roboflow-trained-dataset	58.1
sea-cucumbers-new-tiles	59.8
thermal-cheetah	74.0
tomatoes-2	81.8
trail-camera	72.5
underwater-objects	12.4
varroa-mites-detection–test-set	20.6
wb-prova	47.9
weeds4	53.0
Average	**45.9**

Table 10. **Results on RF100-VL - Flora-Fauna category** for 10-shot Performance (mAP)

with the problem faced by modern deep neural networks [3].

Our second observation is that our proposed strategy achieves the best or the second-best performance with and without OOD samples on all few-shot settings, and the performance reduction is also the lowest compared to the other training configurations. Meanwhile, layer-parallelization on the decoder alone did not significantly alleviate the mAP decline problem. We believe that the diversity brought by

Industrial Datasets	10-shot
-grccs	50.2
13-lkc01	32.6
2024-frc	62.0
aircraft-turnaround-dataset	35.4
asphaltdistressdetection	21.2
cable-damage	24.2
conveyor-t-shirts	38.2
dataconvert	64.8
deeppcb	46.0
defect-detection	52.5
fruitjes	57.1
infraredimageofpowerequipment	47.0
ism-band-packet-detection	60.2
l10ul502	48.5
needle-base-tip-min-max	27.9
recode-waste	39.0
screwdetectclassification	50.2
smd-components	59.0
truck-movement	63.6
tube	67.8
water-meter	48.3
wheel-defect-detection	32.2
Average	**46.7**

Table 11. **Results on RF100-VL - Industrial category** for 10-shot Performance (mAP)

Medical Datasets	10-shot
canalstenosis	46.6
crystal-clean-brain-tumors-mri-dataset	56.1
dentalai	20.4
inbreast	34.2
liver-disease	19.7
nih-xray	8.8
spinefrxnormalvindr	16.6
stomata-cells	17.8
train	7.5
ufba-425	26.6
urine-analysis1	26.0
x-ray-id	46.4
xray	8.1
Average	**25.8**

Table 12. **Results on RF100-VL - Medical category** for 10-shot Performance (mAP)

structural parallelization and the different initialization of each decoder layer gradually weakens with training. However, the input diversity brought by the random initialization of the denoising query can alleviate this problem, allowing the parallel decoder layers to learn different weights, making the final aggregated result more robust and calibrated.

Other Datasets	10-shot
buoy-onboarding	26.1
car-logo-detection	73.9
clashroyalechardetector	40.6
cod-mw-warzone	46.4
countingpills	85.8
everdaynew	50.0
flir-camera-objects	32.2
halo-infinite-angel-videogame	64.4
mahjong	52.1
new-defects-in-wood	32.5
orionproducts	41.9
pill	43.0
soda-bottles	28.8
taco-trash-annotations-in-context	36.8
the-dreidel-project	53.9
Average	**47.2**

Table 13. **Results on RF100-VL - Other category** for 10-shot Performance (mAP)

Finally, we argue that the design of HED, which not only contains the structural parallelization but also introduces the input randomness, can improve the overall prediction accuracy and provide less overconfident results, i.e., more reliable predictions under few-shot settings.

Baseline + Progressive ft. + HED w/ re-init Dn Query	1-shot		5-shot		10-shot	
	w/o OOD	w/ OOD	w/o OOD	w/ OOD	w/o OOD	w/ OOD
ArTaxOr	49.1	44	76.8	73.2	79.2	75.7
DIOR	24.6	23.6	35.3	33.8	41.5	40.3
NEU-DET	15.5	5.3	25.2	14.2	28.4	15.7
UODD	22.1	13.3	27.5	17.0	32.3	25.7
clipart1k	55.6	42.2	59.4	47.5	59.6	44.3
Avg	33.38	**25.68**	44.84	**37.14**	**48.2**	**40.34**
Reduction rate		**-23.07 %**		**-17.17 %**		**-16.31 %**

Baseline + HED w/ re-init Dn Query	1-shot		5-shot		10-shot	
	w/o OOD	w/ OOD	w/o OOD	w/ OOD	w/o OOD	w/ OOD
ArTaxOr	40.5	32.2	71.3	67.5	78.9	74.0
DIOR	22.4	21.0	34.5	32.9	39.3	37.9
NEU-DET	16.4	6.6	24.2	13.0	26.1	13.3
UODD	21.7	14.2	26.3	20.2	32.1	26.1
clipart1k	55.8	40.8	60.0	45.1	60.5	42.4
Avg	31.36	22.96	43.26	35.74	47.38	38.74
Reduction rate		-26.79 %		-17.38 %		-18.24 %

Baseline + Progressive ft. + HED w/o re-init Dn Query	1-shot		5-shot		10-shot	
	w/o OOD	w/ OOD	w/o OOD	w/ OOD	w/o OOD	w/ OOD
ArTaxOr	49.5	42.6	74.7	70.3	79.3	74.0
DIOR	23.9	21.8	34.5	31.7	40.1	36.0
NEU-DET	14.5	5.1	25.5	12.0	26.6	11.0
UODD	22.7	11.4	26.9	17.1	31.0	21.8
clipart1k	57.0	40.5	60.3	44.0	61.1	45.0
Avg	**33.52**	24.28	44.38	35.02	47.62	37.56
Reduction rate		-27.57 %		-21.09 %		-21.13 %

Baseline + Progressive ft.	1-shot		5-shot		10-shot	
	w/o OOD	w/ OOD	w/o OOD	w/ OOD	w/o OOD	w/ OOD
ArTaxOr	44.3	38.8	76.9	73.3	79.1	73.3
DIOR	23.1	21.9	35.3	33.7	42.9	41.0
NEU-DET	14.9	3.6	24.7	7.8	26.5	12.0
UODD	22	14.2	29.1	17.7	30.8	25.0
clipart1k	56.9	40.3	60.3	41.9	60.4	40.9
Avg	32.24	23.76	**45.26**	34.88	47.94	38.44
Reduction rate		-26.3 %		-22.93 %		-19.82 %

Baseline	1-shot		5-shot		10-shot	
	w/o OOD	w/ OOD	w/o OOD	w/ OOD	w/o OOD	w/ OOD
ArTaxOr	38.7	31.4	72.8	66.4	78.0	74.5
DIOR	20.5	19.0	35.1	33.9	41.6	39.7
NEU-DET	11.6	1.3	23.1	10.6	24.9	11.5
UODD	18.9	13.7	26.0	20.0	30.9	24.3
clipart1k	57.5	38.5	60.5	47.4	61.5	40.2
Avg	29.44	20.78	43.5	35.66	47.38	38.04
Reduction rate		-29.42 %		-18.02 %		-19.71 %

Table 14. **Full results on the prediction performance (mAP) on the clean target test set and the proposed CD-Mixed test set, with the performance reduction in percentage.** The best results are highlighted in bold.

References

[1] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. 2

[2] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *ECCV*, 2024. 1

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017. 4

[4] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 3

[5] 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 *ECCV*. Springer, 2014. 3

[6] Longfei Liu, Wen Guo, Shihua Huang, Cheng Li, and Xi Shen. From coco to coco-fp: A deep dive into background false positives for coco detectors. *arXiv preprint arXiv:2409.07907*, 2024. 3

[7] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55. Springer, 2024. 2

[8] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 2

[9] Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-vl: A multi-domain object detection benchmark for vision-language models. *NeurIPS*, 2025. 1, 2

[10] Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. RF-DETR: Neural architecture search for real-time detection transformers. In *ICLR*, 2026. 1

[11] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *NeurIPS*, 2022. 2

[12] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 3