Adapting with an Open Mind: Leveraging Open-Vocabulary Detectors for Closed Set Source-Free Domain Adaptive Object Detection

Supplementary Material

1. Datasets

Our experiments included the usage of multiple datasets, namely Cityscapes[2], Foggy Cityscapes[11], KITTI[3], SIM10k[6], BDD100k[14] and INBreast[10]. The different datasets cover both natural domain images, and medical.

Cityscapes[2] The Cityscapes dataset is a large-scale dataset consisting of stereo video sequences. These sequences were recorded in street scenes from 50 different cities. The dataset has 5000 frames with high quality pixel-level annotations, and an additional larger set with 20000 weakly annotated frames. This work utilizes the training split of $2,975$ annotated images and the remaining 500 annotated images for testing as utilized by prior works using this dataset.

Foggy Cityscapes[11] The Foggy Cityscapes dataset constitutes a collection of synthetic foggy images. A fog simulation pipeline was employed to generate fog on the Cityscapes dataset. Thus, the annotations of the resulting dataset are inherited from the parent dataset, Cityscapes[2]. This work utilizes the training split of $2,975$ annotated images and the remaining 500 annotated images for testing as utilized by prior works using this dataset.

KITTI[3] KITTI is a benchmark dataset used for autonomous driving and mobile robotics. The datasets contains images from real-world street scenes taken from hours of traffic scenarios in rural areas and highways. Up to 15 cars and 30 pedestrians can be found in every image.

SIM10k[6] SIM10k is a synthetic dataset consisting of 10000 images photo-realistic computer images from a simulation engine. Different times of the day, and complex weather and lighting scenarios were simulated spanning across sunny, foggy, rainy and hazy climates during day, night, morning and dusky settings.

BDD100k[14] BDD100k is the largest and most diverse driving video dataset. The videos were recorded from diverse locations across the United States and cover different weather and lighting conditions. This work utilizes the training split of $36,728$ annotated images and the remaining 5258 annotated images for testing as utilized by prior works using this dataset.

INBreast[10] consists of 410 mammography images from 115 patients, comprising of 87 confirmed malignancies. `InBreast` provides digitally acquired mammograms from Portugal providing a resourceful dataset for the adaptation experiments. The domain shift in the medical setting, exists in the difference in the acquisition technologies between the source and target datasets and also the variations in the patient demographics.

Table 1. Experimenting with `GDINO`'s [9] bounding-box threshold

Experimental Setting	AP (car)
Ours + `GDINO` box threshold 0.05	58.7
Ours + `GDINO` threshold 0.35	**70.1**
Ours + `GDINO` threshold 0.75	57.2

Table 2. Weak augmentation methods and their parameter(s)

Augmentation	Parameter(s)
RandomHorizontalFlip	`Probability=0.5`
ResizeImg	`size=800, max_size=1333`
NormalizeImg	`mean=[0.485, 0.456, 0.406]` `std=[0.229, 0.224, 0.225]`

Table 3. Strong augmentation methods and their parameter(s)

Augmentation	Parameter(s)
Color Jittering	`brightness=0.4, contrast=0.4,` `saturation=0.4, hue=0.4`
ResizeImg	`size=800, max_size=1333`
Random Grayscale	`Probability=0.2`
Gaussian Blur	`sigma_min=0.1, sigma_max=2.0` `probability=0.5`
NormalizeImg	`mean=[0.485, 0.456, 0.406]` `std=[0.229, 0.224, 0.225]`

2. Ablation Study (contd.)

We experiment with different bounding box confidence thresholds to filter reliable zero-shot proposals from `GDINO`. As shown in the Tab. 1 for the `S2C` setting, we observe that the threshold of 0.35 provides the best performance and works well for other adaptation settings as well. We observe that lowering the threshold from 0.35 results in false positives and object classes with similarities (such as `person` and `rider`) getting multiple overlapping false positive predictions. On the other hand, increasing the threshold, results in missing out on detection of occluded objects. Hence, empirically, we observe that the chosen threshold provides the best tradeoff of these aspects when extracting the zero-shot predictions from the `OVOD`.

3. Detailed List of Hyperparameters

The strong augmentation transformations used in our adaptation as shown in Table 3 are applied to the input given to the student model and the weak augmentation transformations shown in Table 2 are applied to the input which is given to the teacher and `OVOD GDINO`. The hyperparameters for `CSOD` models `DefDeTR` and `FND` are detailed in Tab. 6 and Tab. 8 respectively. The configuration for `GDINO`

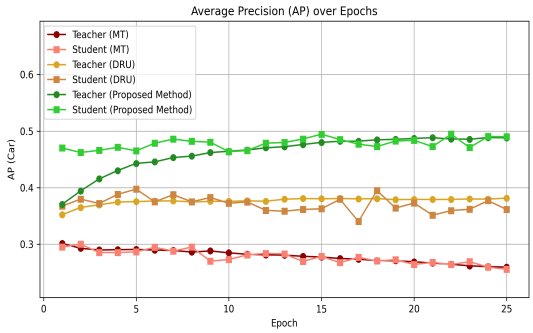

Figure 1. The adaptation graph for the K2C adaptation setting. We show the change in student's and teacher's performance during adaptation when initialized with same source pre-trained weights. The Mean-Teacher degrades after a few epochs. The DRU [7]'s teacher model also saturates eventually. The proposed method anchors the student and ensures that the teacher model gradually improves on the target domain, validating that the quality of pseudo-labels improves with successive training epochs providing an additional advantage in the adaptation process.

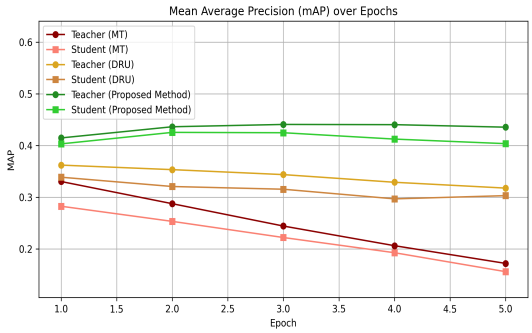

Figure 2. The adaptation graph for the C2B adaptation setting. We show the change in student's and teacher's performance during adaptation when initialized with same source pre-trained weights. The Mean-Teacher degrades after a few epochs. The DRU [7]'s teacher model also saturates eventually. The proposed method anchors the student and ensures that the teacher model gradually improves on the target domain, validating that the quality of pseudo-labels improves with successive training epochs providing an additional advantage in the adaptation process.

for each adaptation setting is detailed in Tab. 7.

4. Dual Branch Decoder Implementation

The dual branch layer design added to the CSOD student model is used to predict two sets of bounding box predictions for a given input. Two identical MLP layers are added to the CSOD model's decoder which project the decoder embeddings into two sets of bounding box predictions. Both MLP layers are initialized with the same set of weights obtained from the source pretrained model. The exact set of hyperparameters for the DefDeTR and FND-based CSOD models are detailed in Tab. 6 and Tab. 8 respectively.

5. Grounding-DINO [9]: Background

GDINO is a popular OVOD for the natural setting which combines the Transformer-based object detector DINO[15] with grounded-pretraining. Along with an image, GDINO takes a text prompt as input which describes the object classes. It effectively fuses the language and images inputs to learn open-set generalizable features to detect new unseen objects guided by the textual prompt. The fusion of vision and languages modalities and its pretraining on several diverse datasets makes it capable of excellent zero-shot object detection in the Natural domain. GDINO achieves an impressive 52.5 AP on the COCO[8] benchmark. In this study, we showed that the domain-invariant representations learnt by OVOD models such as GDINO can be used to remove the source domain bias in the CSOD model. Figure 1 and 2 show the training graphs for the K2C and C2B settings respectively demonstrating that the mitigation of source-domain bias in the pseudo labels results in higher quality pseudo label in the successive epochs of the adaptation.

6. BiomedParse [16]: Background

BMP is another OVOD model, designed for comprehensive biomedical image analysis. BMP makes use of the modular SEEM [18] architecture, using Focalnet [13] as its image encoder and PubMedBERT [4] as the text encoder, and provides a unified approach to perform joint Segmentation, Detection and Recognition across 9 modalities of diverse biomedical images. BMP was trained on over 6 million triples of image, segmentation mask and textual description, from over 1 million images, with labels spanning 82 major biomedical object types across 9 imaging modalities. During inference, BMP takes a text prompt, along with an image. In our proposed methodology, we demonstrate how a extremely challenging Breast cancer detection task by adapting a model to the publicly available InBreast dataset. The text prompt used for BMP is *"Its margins can show small undulations or lobulations"*. The text prompt was created using the assistance of an expert radiologist with over 15 years of clinical experience.

7. Deformable-DeTR [17]: Background

Deformable DETR (DefDeTR) extends the original DeTR architecture by addressing its two major limitations i.e slow training convergence and suboptimal performance on small objects. DefDeTR introduces multi-scale deformable attention, a sparse attention mechanism that samples a limited set of key points around reference anchors, enabling efficient aggregation of information across feature pyramid levels. In its standard configuration, DefDeTR

Table 4. We also extend our validation on challenging medical imaging task of breast cancer detection from mammograms where the source-model is pretrained on DDSM [12] and adapted to the publicly available InBreast [10].

Method	Venue	SF	R@0.3	R@0.5
D-MASTER [1]	MICCAI'24	✗	0.60	0.72
DRU [7]	ECCV'24	✓	0.57	0.64
DDT[5]	ICCV'25	✓	0.61	0.68
Ours (DefDeTR)		✓	**0.70**	**0.76**

adopts a ResNet-50 backbone to generate rich multi-scale feature maps. This choice offers a strong trade-off between representational capacity and efficiency, making it suitable for training on large-scale datasets without excessive computational overhead. The backbone outputs multiple resolution stages that naturally align with the multi-scale design of the deformable attention module, enhancing localization and robustness. Although ResNet-50 is the most commonly used backbone in benchmark experiments and official implementations, DefDeTR is flexible and can be equipped with a range of stronger or lighter feature extractors.

8. FocalNet-DINO [13]: Background

FocalNet-DINO [13](FND) integrates the dense self-attention capabilities of the DINO [15] detector with the efficient hierarchical representations of FocalNet, a transformer backbone designed to enhance fine-grained feature modeling. FocalNet introduces focal modulation, a lightweight alternative to multi-head self-attention that aggregates both local and global context using modulated receptive fields, enabling the backbone to capture rich spatial relationships at multiple scales with significantly reduced computational cost. When paired with the DINO detection head, which employs denoising training, mixed query selection, and contrastive matching losses, FocalNet-DINO yields a highly stable and robust training pipeline that improves convergence and enhances performance on complex detection tasks. The FocalNet-T and FocalNet-S variants are commonly used, but the detector is also compatible with larger FocalNet-B/L models for higher-capacity settings.

9. Adaptation on public mammography datasets

We present another example of adaptation using our method from the public mammography dataset DDSM [12] to InBreast [10]. DDSM is one of the earliest mammography datasets and consists of digitized versions of film-based mammograms, whereas InBreast contains mammograms acquired directly using digital systems. This setting exhibits a substantial domain shift, as the appearance of mammograms in DDSM and InBreast differs considerably due to the distinct acquisition technologies. For this experiment, we use the publicly released D-MASTER [1] weights and initialize all models reported in Tab. 4 accordingly. Despite the challenging nature of this adaptation scenario, our method achieves a notable 9% improvement in recall at a strict threshold of 0.3 false positives per image (FPI), and an 8% improvement in recall at 0.5 FPI over DDT [5].

Table 5. Curriculum hyperparameters for each adaptation setting

Setting	Total epochs	T_0
C2F [DefDeTR]	60	30
C2B [DefDeTR]	10	5
S2C [DefDeTR]	60	35
K2C [DefDeTR]	30	5
C2F [Focal-Net DINO]	30	2
C2B [Focal-Net DINO]	5	1
S2C [Focal-Net DINO]	15	5
K2C [Focal-Net DINO]	15	5
Private2Inbreast	80	1
DDSM2Inbreast	80	1

10. Configuration of the training curriculum

Tab. 5 summarizes the curriculum hyperparameters used for each adaptation setting. As defined in the main paper, T_0 denotes the delay (in number of epochs) after which the loss term $\mathcal{L}_{\text{OVOD}}$ begins to contribute. The values of T_0 are selected empirically, allowing the model to first learn domain-specific features for T_0 epochs before incorporating $\mathcal{L}_{\text{OVOD}}$. We observe that for FND, introducing $\mathcal{L}_{\text{OVOD}}$ earlier is beneficial. Since FND is a stronger detector, it requires fewer epochs to acquire domain-specific representations, enabling the loss to be applied sooner as compared to DefDeTR. In contrast, for the medical setting, delaying the introduction of $\mathcal{L}_{\text{OVOD}}$ is not advantageous. The adaptation process is highly sensitive to source domain bias, and therefore we find it preferable to apply $\mathcal{L}_{\text{OVOD}}$ from the beginning of training.

11. Additional Visualizations

Fig. 3 shows some additional visualizations on the C2F, C2B, K2C dataset. It can be observed that Mean-Teacher often cannot overcome the source-domain biases on its own and this can result in weak and suboptimal adaptation to the target domain. On the other hand, G-DINO (OVOD) stays robust to different levels of domain shift but may not be as good as a close set object detector (CSOD). As seen in Fig. 3, our method is successful is leveraging the robustness to domain shift from the OVOD and distill it to the underlying CSOD resulting in an overall improvement to the target domain adaptation for both DefDeTR and FND-based CSOD models.

Table 6. Detailed list of Hyperparameters for the Deformable-DeTR-based student and teacher models

Hyperparameter	Value	Description
backbone	resnet50	Backbone used for Deformable-DeTR
pos_encoding	sine	Sinusoidal positional encoding for the input tokens
num_classes	9	Number of object classes
num_queries	300	Total number of object queries
num_feature_levels	4	Number of feature levels
with_box_refine	store_true	Enabling the box refine function
hidden_dim	256	MLP hidden layer dimension
num_heads	8	Number of output heads
num_encoder_layers	6	Total number of encoder layers
num_decoder_layers	6	Total number of decoder layers
feedforward_dim	1024	Dimension of the feed-forward network.
dropout	0.0	Dropout
batch_size	8	Training batch size
eval_batch_size	8	Evalutaion batch size
lr	2e-4	Learning rate
lr_backbone	2e-5	Learning rate for the backbone
lr_linear_proj	2e-5	Learning rate for linear projection layer
weight_decay	1e-4	Rate of weight decay
clip_max_norm	0.5	Gradient clipping max norm
epoch	30	Default number of epochs for training
coef_class	2.0	Weight coefficient for the cross-entropy loss in $\mathcal{L}_{\text{CSOD}}$ and $\mathcal{L}_{\text{OVOD}}$
coef_boxes	5.0	Weight coefficient for the bounding-box loss in $\mathcal{L}_{\text{CSOD}}$ and $\mathcal{L}_{\text{OVOD}}$
coef_giou	2.0	Weight coefficient for the GIOU loss in $\mathcal{L}_{\text{CSOD}}$ and $\mathcal{L}_{\text{OVOD}}$
alpha_ema	0.999	EMA coefficient
masked_ratio	0.5	Masking ratio for augmentation
coef_masked_img	1.0	Coefficient of masked image
threshold	0.3	Confidence threshold for filtering the pseudo-labels.
expert_model	groundingdino	Grounding-DINO as the OVOD model.

Table 7. Detailed list of Hyperparameters for GDINO [9]

Hyperparameter	Setting	Value	Description
grounding_dino_config	C2F	GroundingDINO_SwinT_OGC.py	Publicly available Grounding-DINO with Swin-T backbone
text_prompt	C2F	"person . car . train . rider . truck . motorcycle . bicycle . bus ."	Text prompt consisting of the required object classes
label_classes	C2F	person,car,train,rider,truck,motorcycle,bicycle,bus	Set of target object classes.
box_threshold	C2F	0.35	Confidence threshold for filtering Grounding-DINO BBoxes.
text_threshold	C2F	0.25	Textual threshold for image-text similarity in Grounding-DINO
num_classes	C2F	9	Number of object classes
grounding_dino_config	C2B	GroundingDINO_SwinT_OGC.py	Publicly available Grounding-DINO with Swin-T backbone
text_prompt	C2B	"person . car . train . rider . truck . motorcycle . bicycle . bus ."	Text prompt consisting of the required object classes
label_classes	C2B	person,car,train,rider,truck,motorcycle,bicycle,bus	Set of target object classes.
box_threshold	C2B	0.35	Confidence threshold for filtering Grounding-DINO BBoxes.
text_threshold	C2B	0.25	Textual threshold for image-text similarity in Grounding-DINO
num_classes	C2B	9	Number of object classes
grounding_dino_config	S2C	GroundingDINO_SwinT_OGC.py	Publicly available Grounding-DINO with Swin-T backbone
text_prompt	S2C	"car ."	Text prompt consisting of the required object classes
label_classes	S2C	car	Set of target object classes.
box_threshold	S2C	0.35	Confidence threshold for filtering Grounding-DINO BBoxes.
text_threshold	S2C	0.25	Textual threshold for image-text similarity in Grounding-DINO
num_classes	S2C	2	Number of object classes
grounding_dino_config	K2C	GroundingDINO_SwinT_OGC.py	Publicly available Grounding-DINO with Swin-T backbone
text_prompt	K2C	"car ."	Text prompt consisting of the required object classes
label_classes	K2C	car	Set of target object classes.
box_threshold	K2C	0.35	Confidence threshold for filtering Grounding-DINO BBoxes.
text_threshold	K2C	0.25	Textual threshold for image-text similarity in Grounding-DINO
num_classes	K2C	2	Number of object classes

Table 8. Detailed list of Hyperparameters for the FND-based student and teacher models

Hyperparameter	Value	Description
num_classes	Number of classes	Number of object classes.
lr	0.0001	Learning rate.
lr_backbone_names	['backbone.0']	Layers with separate LR.
lr_linear_proj_names	['reference_points','sampling_offsets']	Layers using LR multiplier.
lr_linear_proj_mult	0.1	LR multiplier for projection layers.
ddetr_lr_param	False	Deformable DETR LR handling.
batch_size	4	Batch size.
weight_decay	0.0001	Weight decay.
epochs	15	Total training epochs.
lr_drop	11	LR drop epoch.
save_checkpoint_interval	1	Save checkpoint every N epochs.
clip_max_norm	0.1	Gradient clipping norm.
onecyclelr	False	Enable OneCycle LR.
multi_step_lr	False	Use multi-step LR.
lr_drop_list	[33, 45]	Epochs for LR drops.
modelname	'dino'	Detector model name.
frozen_weights	None	Optional pretrained weights.
backbone	'focalnet_L_384_22k'	Backbone architecture.
focal_levels	3	FocalNet levels.
focal_windows	5	Focal window size.
use_checkpoint	True	Gradient checkpointing.
lr_backbone	1e-05	Backbone LR.
dilation	False	Use dilated convolutions.
position_embedding	'sine'	Type of positional encoding.
backbone_freeze_keywords	None	Layers to freeze.
enc_layers	6	Encoder layers.
dec_layers	6	Decoder layers.
unic_layers	0	UniC layers.
pre_norm	False	Pre-normalization flag.
dim_feedforward	2048	Transformer FFN dim.
hidden_dim	256	Hidden dimension.
dropout	0.0	Dropout rate.
nheads	8	Attention heads.
num_queries	900	Number of queries.
query_dim	4	Query dimension.
num_patterns	0	Pattern embeddings.
aux_loss	True	Auxiliary losses.
set_cost_class	2.0	Class matching cost.
set_cost_bbox	5.0	Box matching cost.
set_cost_giou	2.0	GIoU matching cost.
cls_loss_coef	1.0	Classification loss weight.
mask_loss_coef	1.0	Mask loss weight.
dice_loss_coef	1.0	Dice loss weight.
bbox_loss_coef	5.0	Bbox loss weight.
giou_loss_coef	2.0	GIoU loss weight.
enc_loss_coef	1.0	Encoder loss weight.
interm_loss_coef	1.0	Intermediate loss weight.
no_interm_box_loss	False	Disable intermediate bbox loss.
focal_alpha	0.25	Focal loss alpha.
use_dn	True	Use denoising queries.
dn_number	100	Number of DN queries.
dn_box_noise_scale	0.4	Noise for DN boxes.
dn_label_noise_ratio	0.5	Noise ratio for DN labels.
embed_init_tgt	True	Init targets in embedding.
dn_labelbook_size	Number of classes + 1	DN labelbook size .

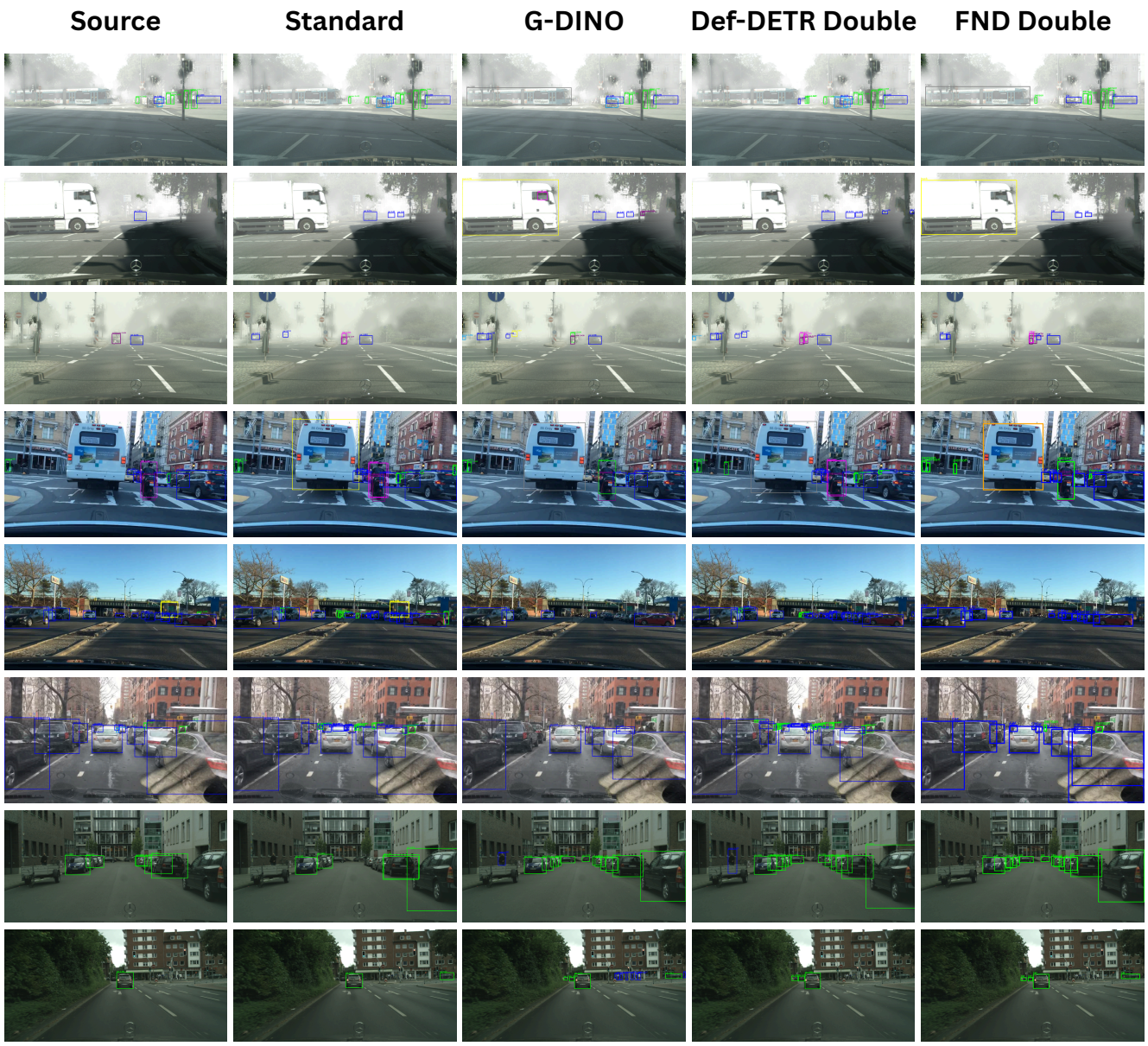

Figure 3. This figures shows the qualitative results for each of the models. Column 1 corresponds to the predictions visualized for the source-pretrained model applied on the target dataset without any adaptation. Column 2 corresponds to the Standard Mean-Teacher. Column 3 corresponds to the zero-shot prediction from `GDINO`, Column 4 and 5 correspond to the predictions using our proposed method using the `DefDeTR` and the `FND` base detectors respectively.

References

[1] Tajamul Ashraf, Krithika Rangarajan, Mohit Gambhir, Richa Gauba, and Chetan Arora. D-MASTER: Mask Annealed Transformer for Unsupervised Domain Adaptation in Breast Cancer Detection from Mammograms . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland, 2024. 3

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[4] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021. 2

[5] Qi He, Xiao Wu, Jun-Yan He, and Shuai Li. Dual-rate dynamic teacher for source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2067–2076, 2025. 3

[6] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 1

[7] Trinh Le Ba Khanh, Huy-Hung Nguyen, Long Hoang Pham, Duong Nguyen-Ngoc Tran, and Jae Wook Jeon. Dynamic retraining-updating mean teacher for source-free object detection. *arXiv preprint arXiv:2407.16497*, 2024. 2, 3

[8] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[9] 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. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2, 4

[10] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 1, 3

[11] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1

[12] Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). *(No Title)*, 2016. 3

[13] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 2, 3

[14] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1

[15] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 3

[16] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024. 2

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*. OpenReview.net, 2021. 2

[18] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 2