TITAN: Query-Token based Domain Adaptive Adversarial Learning

Supplementary Material

***Sec.* A** *provides additional theoretical insights.*

***Sec.* B** *provides additional details on training datasets.*

***Sec.* C** *provides additional details on hyperparamers.*

***Sec.* C** *provides further architectural details.*

***Sec.* D** *provides details about multiple backbones.*

***Sec.* E** *provides additional details on augmentations.*

***Sec.* F** *provides extensive ablation and visualizations.*

A. Detailed Theoretical Analysis

Proof of Preposition. We denote the spaces of the output functions $F_{(A_1,\dots,A_{i-1})}$ induced by the weight matrices $A_i, i = 1, \dots, 5$ by $\mathcal{H}_i, i = 1, \dots, 5$, respectively. Lemma A.7 in [2] suggests the following inequality:

$$\begin{aligned}\log \mathcal{N}(\mathcal{F}|S) &\leq \log \left(\prod_{i=1}^{5} \sup_{\mathbf{A}_{i-1} \in \boldsymbol{\mathcal{B}}_{i-1}} \mathcal{N}_i \right) \\ &\leq \sum_{i=1}^{5} \log \left(\sup_{\substack{(A_1,\dots,A_{i-1}) \\ \forall j<i, A_j \in B_j}} \mathcal{N}\left(\left\{A_i F_{(A_1,\dots,A_{i-1})}\right\}, \varepsilon_i, \|\cdot\|_2\right) \right)\end{aligned} \tag{1}$$

Thus, we obtain the following inequality:

$$\log \mathcal{N}(\mathcal{F}|S) \leq \sum_{i=1}^{5} \frac{b_i^2 \|F_{(A_1,\dots,A_{i-1})}(X)\|_2^2}{\varepsilon_i^2} \log\left(2W^2\right) . \tag{2}$$

Meanwhile, we have the following inequality:

$$\begin{aligned}\|F_{(A_1,\dots,A_{i-1})}(X)\|_2^2 &= \|\sigma_{i-1}(A_{i-1}F_{(A_1,\dots,A_{i-2})}(X)) - \sigma_{i-1}(0)\|_2 \\ &\leq \|\sigma_{i-1}\| \|A_{i-1}F_{(A_1,\dots,A_{i-2})}(X) - 0\|_2 \\ &\leq \rho_{i-1}\|A_{i-1}\|_\sigma \|F_{(A_1,\dots,A_{i-2})}(X)\|_2 \\ &\leq \rho_{i-1}s_{i-1}\|F_{(A_1,\dots,A_{i-2})}(X)\|_2.\end{aligned} \tag{3}$$

Therefore, we get:

$$\|F_{(A_1,\dots,A_{i-1})}(X)\|_2^2 \leq \|X\|^2 \prod_{j=1}^{i-1} s_i^2 \rho_i^2 . \tag{5}$$

Motivated by the proof in [2], we assume the following equations:

$$\begin{aligned}\varepsilon_{i+1} &= \rho_i s_{i+1} \varepsilon_i \\ \varepsilon_5 &= \rho_1 \prod_{i=2}^{4} s_i \rho_i s_5 \epsilon_1 \\ \varepsilon &= \rho_1 \prod_{i=2}^{5} s_i \rho_i \epsilon_1\end{aligned} \tag{6}$$

Therefore, we have:

$$\varepsilon_i = \frac{\rho_i \prod_{j=1}^{i-1} s_j \rho_j}{\prod_{j=1}^{5} s_j \rho_j} \varepsilon . \tag{7}$$

Thus, we obtain:

$$\log \mathcal{N}(\mathcal{F}|_S, \varepsilon, \|\cdot\|_2) \leq \frac{\log\left(2W^2\right) \|X\|_2^2}{\varepsilon^2} \left(\prod_{i=1}^{5} s_i \rho_i \right)^2 \sum_{i=1}^{5} \frac{b_i^2}{s_i^2} \tag{8}$$

which is precisely Equation (13). The proof is completed. □

B. Additional Details on Training datasets

Our experiments are conducted on five `SF-DAOD` benchmark datasets. Along with this, we introduce the `SF-DAOD` problem in the medical domain and conduct experiments on four Breast Cancer Detection (`BCD`) datasets, including two publicly available datasets.

B.1. Natural Datasets

1) **Cityscapes** [15] is gathered from urban environments across 50 European cities, provides detailed annotations for 30 semantic classes across 8 categories. It comprises 5,000 high-quality annotated images and a larger set of 20,000 coarsely annotated images, all high-resolution (2048x1024 pixels) providing detailed visual data for precise scene understanding. Our study utilizes the high-quality subset of 5,000 images, consisting of 2,975 training images and a standard test set of 500 images from Frankfurt, Munster, and Lindau, as employed in prior research. This dataset is release under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, and is available for academic and research purposes.

2) **Foggy Cityscapes** [82] is an extension of the Cityscapes dataset, designed to support research in developing robust computer vision algorithms for autonomous

Table 6. Details of Training Datasets

Dataset	Type	Pre-training (Source)		Unsupervised Adaptation (Target)	
		Train	Val/Test	Train	Test
Cityscapes	Natural	2,975	500	2,975	500
Foggy Cityscapes	Natural	-	-	2,975	500
KITTI	Natural	7,481		-	-
SIM10k	Natural	8,500	1,500	-	-
BDD100k	Natural	-	-	36,728	5,258
RSNA-BSD1K	Medical	1000 (200)	250 (50)	1000 (200)	1000(200)
INBreast	Medical	-	-	410 (91)	410 (91)
DDSM	Medical	2885 (1339)	218 (118)	3103 (1458)	3103 (1458)

driving in foggy conditions. Similarly, it consists of high-resolution images (2048x1024 pixels) with annotations ofr 2D bounding boxes, pixel-level semantic segmentation, and instance segmentation, inherited from the original dataset. It is directly constructed from Cityscapes by simulating three levels of foggy weather (0.005,0.001,0.02), but we adapt on the most extreme level (0.02) in our experiments as done by [82]. This dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3) **KITTI** [27] is a promenient benchmark for computer vision and robotics, particularly in autonomous driving. Data was collected using a car-mounted sensor suite, including high-resolution color and grayscale cameras, a Velodyne laser scanner, and a GPS/IMU system, in urban, rural, and highway settings around Karlsruhe, Germany. The dataset includes stereo image pairs with disparity maps, consecutive frames for optical flow, image sequences with ground truth poses for SLAM, and images with corresponding 3D point clouds for object detection. The images are typically high-resolution (1242x375 pixels), captured at 10-100 Hz, providing detailed and diverse visual data. The 3D object detection subset is split into 7,481 training images and 7,518 test images. As quite common in the Domain Adaptation literature, we only use the 7,481 training images in our experiments for adaptation and evaluate on the same split as done in . The dataset is publicly available for academic use under the Creative Commons Attribution-NonCommercial-ShareAlike license.

4) **SIM10k** [46] is a synthetic dataset create using the Grand Theft Auto V (GTA V) engine, it simulates various driving scenarios providing diverse set of high-resolution images with detailed annotations. This dataset consists of 10,0000 images of urban environments under different weather conditions, lighting and traffic situations. In our experiments for source training, we prepare our own train and val set of 8,500 and 1,500 images respectively. We intend to make this split public for reproducibility. This dataset is also publicly available for academic and research purposes, with usage terms typically provided by the dataset creators, allowing for non-commercial use.

5) **BDD100k** [105] developed by the Berkeley DeepDrive (BDD) team, is one of the largest and most diverse driving video datasets available. It comprises of 100,000 driving videos and 100K keyframe images, and captures a wide range of driving scenarios across urban, suburban and rural environments in the United States, under diverse weather conditions, lighting, and times of day. Each image is of 720p and is richly annotated with 2D bounding boxes for objects, lane markings, drivable areas and scene attributes like weather and time of day. We make use of the standard BDD100K train set containing 36,728 images for adaptation and use 5,258 images for evaluation. These images are the frames at the 10th second in the videos and the split is consistent with the original video set. This dataset is released under the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 Internation license and is publicly available for academic and research purposes.

B.2. Medical Datasets

INBreast [69] is a relatively small breast cancer detection dataset, consisting of 410 mammography images from 115 patients, including 87 confirmed malignancies. The training images include both histologically confirmed cancers and benign lesions initially recalled for further examination but later identified as nonmalignant. We hypothesize that incorporating both malignant and benign lesions in the training process will enhance our model's ability to detect a broader range of lesions and effectively distinguish between malignant and benign cases.

DDSM [51] is a publicly available breast cancer detection dataset, comprising 2,620 full mammography images, with 1,162 containing malignancies. The DDSM dataset offers digitized film-screen mammography exams with lesion annotations at the pixel level, where cancerous lesions are histologically confirmed. We used the DDSM dataset exclusively for training our model and not for evaluation. This decision stems from the observation that the quality of digitized film-screen mammograms is inferior to that of full-field digital mammograms, making evaluation on these cases less relevant. For our purposes, we converted the lossless JPEG images to PNG format, mapped pixel values to optical density using calibration functions from the DDSM website, and rescaled the pixel values to a range of 0–255.

RSNA-BSD1K [7] is a comprehensive collection of 54,706 screening mammograms sourced from approximately 8,000 patients. This dataset includes a diverse range of cases, among which 1,000 instances have been identified as malignant. The dataset serves as a valuable resource for developing and evaluating machine learning models in the field of medical imaging, particularly in breast cancer detection. From this large dataset, a specialized subset is curated known as RSNA-BSD1K, which consists of 1,000 carefully selected mammograms. This subset was designed to maintain a balance between normal and malignant cases while ensuring high-quality annotations suitable for robust model training and evaluation. Within RSNA-BSD1K, 200 cases have been confirmed as malignant, representing a diverse spectrum of tumor characteristics and imaging conditions.

Note that unlike in natural images, single domain detection techniques which use a particular subset of the dataset for adaptation and remaining for testing, our technique does not require any labels from the target dataset. Hence, for medical datasets, it seems logical to use the whole dataset during training and testing, and not just the any train or test split. Hence, when reporting results for "Dataset A to Dataset B", we imply that the model is trained on $\mathcal{D}_s = A$ (whole dataset for the training), and adapted for $\mathcal{D}_t = B$ (whole dataset for adaptation in an unsupervised way and testing). Table X shows the detailed split wise sets used during experiments.

C. Further Insights into Hyperparameter Selection

The selection of hyperparameters, as detailed in Table 7, was guided by empirical experimentation and domain-specific considerations. Key factors included optimizing model generalization, ensuring stability during training, and balancing performance across different benchmarks. Parameters such as the number of pseudo labels, learning rate, and loss weights were fine-tuned based on validation results, with adjustments made dynamically for specific dataset shifts. Additionally, threshold values for pseudo-labeling were set adaptively to enhance robustness across diverse datasets.

Table 7. Below are the detailed hyper-parameters corresponding to each benchmark, with the source dataset as the In-house dataset

Hyper-parameter	Description	Value
num_classes	Number of classes	1
lr	Learning rate	0.0001
lr_backbone	Learning rate for backbone	1e-05
batch_size	Batch size	4
weight_decay	Weight decay	0.0001
epochs	Number of epochs	100
lr_drop	Learning rate drop	11
clip_max_norm	Clip max norm	0.1
multi_step_lr	Multi-step learning rate	True
modelname	Model name	'dino'
backbone	Backbone	'focalnet_L_384_22k_fl4'
focal_levels	Focal levels	4
focal_windows	Focal windows	3
position_embedding	Position embedding	'sine'
pe_temperature	PE temperature	20
enc_layers	Encoder layers	6
dec_layers	Decoder layers	6
dim_feedforward	Dimension of feedforward network	2048
hidden_dim	Hidden dimension	256
dropout	Dropout	0.0
nheads	Number of heads	8
num_queries	Number of queries	900
box_attn_type	Box attention type	'roi_align'
num_feature_levels	Number of feature levels	4
enc_n_points	Encoder points	4
dec_n_points	Decoder points	4
transformer_activation	Transformer activation	'relu'
batch_norm_type	Batch norm type	'FrozenBatchNorm2d'
set_cost_class	Set cost class	2.0
set_cost_bbox	Set cost bbox	5.0
set_cost_giou	Set cost GIoU	2.0
cls_loss_coef	Class loss coefficient	1.0
mask_loss_coef	Mask loss coefficient	1.0
dice_loss_coef	Dice loss coefficient	1.0
bbox_loss_coef	BBox loss coefficient	5.0
giou_loss_coef	GIoU loss coefficient	2.0
enc_loss_coef	Encoder loss coefficient	1.0
focal_alpha	Focal alpha	0.25
matcher_type	Matcher type	'HungarianMatcher'
nms_iou_threshold	NMS IoU threshold	0.1
use_dn	Use DN	True
dn_number	DN number	100
dn_box_noise_scale	DN box noise scale	1.0
dn_label_noise_ratio	DN label noise ratio	0.5
dn_labelbook_size	DN labelbook size	3
use_ema	Use EMA	False
ema_decay	EMA decay	0.9997
optim_iter_per_epoch	Optimization iterations per epoch	2500

D. More details on different Backbones

RN50_4scale and RN50_5scale. ResNet-50 is a 50-layer deep convolutional neural network designed for image recognition tasks. It employs residual learning to allow the network to learn residual functions with reference to the layer inputs, which helps in training deeper networks. The model is pretrained on the ImageNet-1k dataset, which contains 1.2 million images and 1,000 classes. This pretraining helps the network to learn robust feature representations that can be fine-tuned for various downstream tasks. Since some methods adopt 5 scales of feature maps and some adopt 4, we report our results with both 4 and 5 scales of feature maps.

Table 8. Detailed hyper-parameters corresponding to each benchmark, with the source dataset as the In-house dataset.

Hyper-parameter	Description	City2Foggy	City2BDD	Sim2City	Kitti2City	InH2InB	InH2DDSM
num_classes	Number of classes	8	2	2	2	2	2
epochs	Number of epochs	10	10	10	10	20	20
topk_pseudo	Number of pseudo labels	30	30	30	30	15	15
use_dynamic_th	Use dynamic threshold	True	True	True	True	False	False
pseudo_th	Initial pseudo label threshold	0.006	0.04	0.04	0.04	0.06	0.06
lambda	Weight of Enc and Dec loss	0.6	0.6	0.6	0.6	1.0	1.0

Table 9. Data Augmentation Methods and Parameters

Augmentation Type	Method	Parameters
Weak	Random Horizontal Flip	Probability = 0.5
	Resize	Size = 800, Max Size = 1333
Strong	Color Jitter (Color Adjustment)	Brightness = 0.4
		Contrast = 0.4
		Saturation = 0.4
		Hue = 0.1
	Random Grayscale	Probability = 0.2
	Gaussian Blur	Sigma Range = [0.1, 2.0
		Probability = 0.5
	Normalization	Mean = [0.485, 0.456, 0.406]
		Std = [0.229, 0.224, 0.225]

Convnext This is a modern convolutional network architecture that aims to bridge the gap between convolutional networks and vision transformers. ConvNeXt incorporates design principles from transformers, such as a simplified architecture with fewer layers and parameters, but retains the efficiency and scalability of convolutional networks. It leverages state-of-the-art techniques like LayerScale, deep supervision, and various normalization methods to achieve competitive performance on benchmarks. ConvNeXt models are often pretrained on large datasets like ImageNet-1k or ImageNet-22k to provide strong initial weights for transfer learning.

Swin The Swin Transformer, particularly the large version (SwinL), is a hierarchical vision transformer designed for image classification, object detection, and segmentation tasks. It introduces the concept of shifted windows for computing self-attention, which allows it to handle varying scales of features more efficiently than traditional transformers. SwinL is pretrained on the ImageNet-22k dataset, which contains 14 million images and 22,000 classes. This extensive pretraining helps the model to capture rich and diverse feature representations, making it highly effective for a wide range of visual tasks. **FN-Fl3 and FN-Fl4** At the core of Focal Modulation Networks (FocalNets) is the focal modulation mechanism: A lightweight element-wise multiplication as the focusing operator to allow the model to see or interact with the input using the proposed modulator; As depicted below, the modulator is computed with a focal aggregation procedure in two steps: focal contextualization to extract contexts from local to global ranges at different levels of granularity and gated aggregation to condense all context features at different granularity levels into the modulator. We adopt the same stage configurations and hidden dimensions as those used in Focal Transformers, but we replace the Self-Attention (SA) modules with Focal Modulation modules. This allows us to construct various Focal Modulation Network (FocalNet) variants. In FocalNets, we only need to define the number of focal levels (L) and the kernel size (k) at each level. For simplicity, we increase the kernel size by 2 for each subsequent focal level, i.e., $k_l = k_{l-1} + 2$. To match the complexities of Focal Transformers, we design both small receptive field (SRF) and large receptive field (LRF) versions for each of the four layouts by using 3 and 4 focal levels, respectively. We use non-overlapping convolution layers for patch embedding at the beginning (with a kernel size of 4 × 4 and stride of 4) and between stages (with a kernel size of 2 × 2 and stride of 2).

The results of utilizing each backbone in our model and the corresponding results are present in Table X.

E. More Details on Augmentations

In our study, we employed both weak and strong augmentation techniques to enhance the robustness and generalization capabilities of our model. These augmentations are applied to the training images to simulate various real-world scenarios and improve the model's performance.

Weak Augmentation

Weak augmentations are relatively simple transformations that slightly alter the images without significantly changing their content. We utilized two primary weak augmentation techniques. The first is *Random Horizontal Flip*, which flips the image horizontally with a probability of 0.5. This helps the model learn invariance to the left-right orientation of objects, making it more robust to such variations. The second technique is *Resize*, where images are resized such that the shortest edge is 800 pixels while maintaining the aspect ratio, and if the longest edge exceeds 1333 pixels, the image is scaled down accordingly. This resizing standardizes the input image size, ensuring consistent and efficient training.

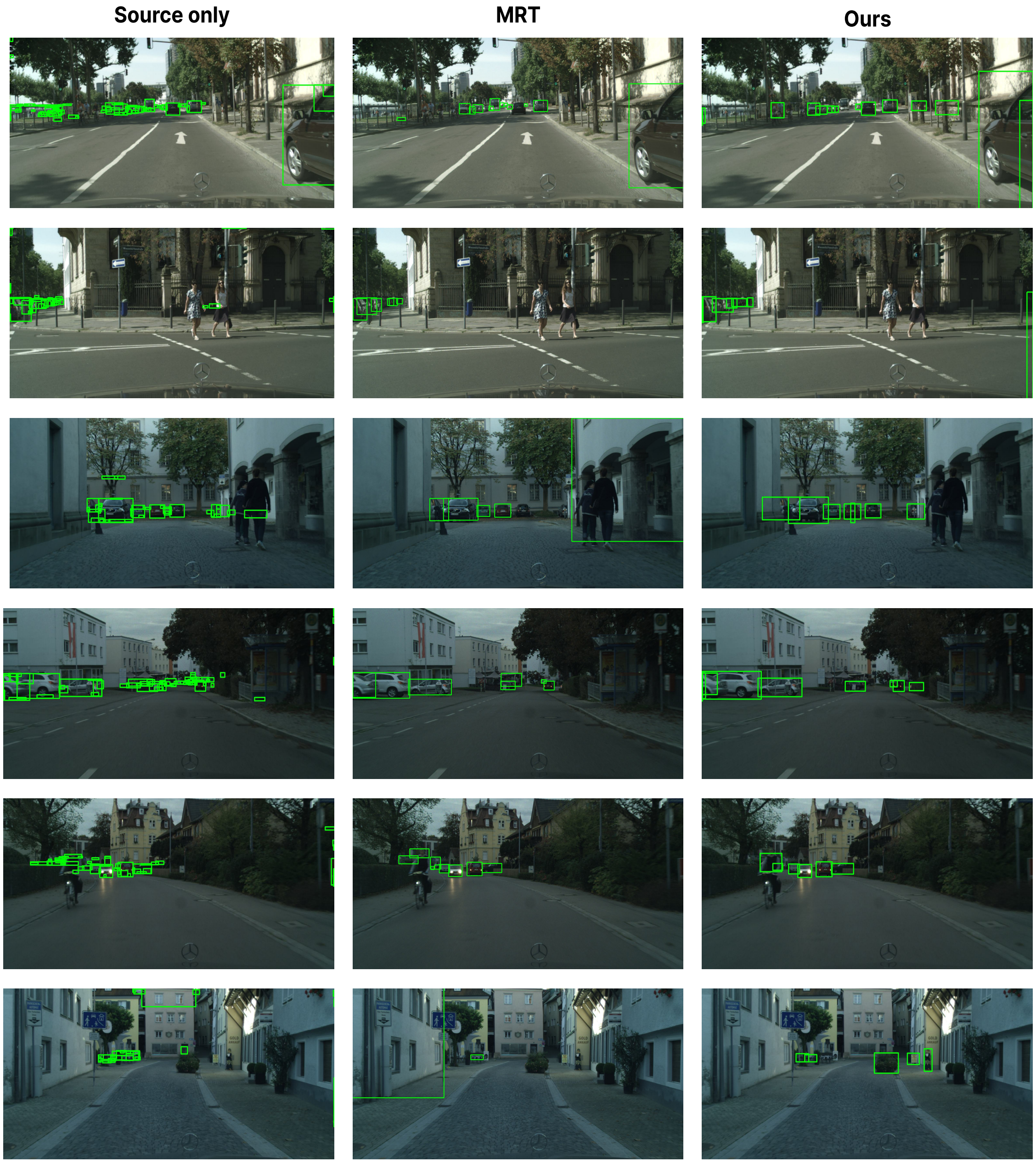

Figure 5. Qualitative results for car detection on Cityscapes in `S2C` setting. `MRT` is an Unsupervised domain adaptation technique with the best results in `S2C`. Our method, being an `SF-DAOD` technique, comes surprisingly close to the best performing `UDA` method.

Together, these weak augmentations provide a baseline level of variability in the training data, helping the model to generalize better across different image scales and orientations.

Strong Augmentation

Strong augmentations involve more complex transformations that significantly alter the images, thereby providing a broader range of variability. These augmentations are designed to challenge the model and improve its ability to handle diverse and complex real-world scenarios. The first strong augmentation is *Color Jitter*, which randomly changes the brightness, contrast, saturation, and hue of the images with specified parameters (brightness = 0.4, contrast = 0.4, saturation = 0.4, hue = 0.1) and a probability of 0.8. This helps the model become invariant to different lighting conditions and color variations. The second augmentation is *Random Grayscale*, which converts the image to grayscale with a probability of 0.2, encouraging the model to focus on shapes and structures rather than colors. The third strong augmentation is *Gaussian Blur*, applied with a sigma range of [0.1, 2.0] and a probability of 0.5, simulating out-of-focus conditions and reducing high-frequency noise. Finally, all images are converted to tensors and normalized using the mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. These strong augmentations introduce substantial variability in the training data, forcing the model to learn more robust and generalized features.

F. Qualitative Analysis

We provide a qualitative visualization analysis of pseudo-labels and feature distributions. Our method outperforms the state-of-the-art methods, including MRT [113] and GT. Figure 7 and Fig. 6 present predictions on the breast cancer dataset and foggy dataset respectively, highlighting the superior performance of our approach.

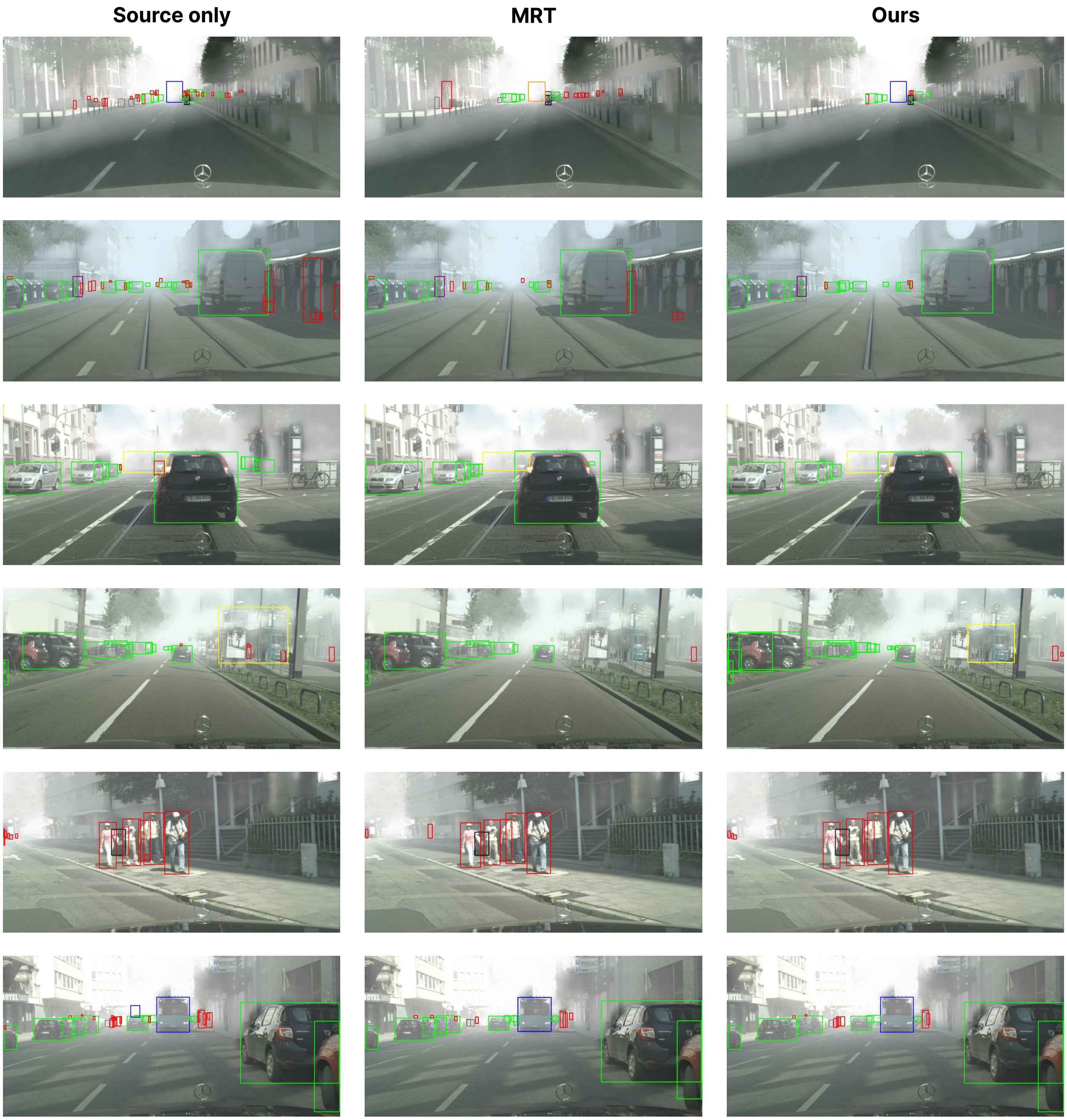

Figure 6. Qualitative results for car detection on Foggy Cityscapes in C2F setting. The visible detections are from the classes: Person, Car, Train, Bicycle, Bus, Truck, motorcycle, Rider. MRT is an Unsupervised domain adaptation technique with the best results in C2F. Our method, being an SF-DAOD technique, comes surprisingly close to the best performing UDA method.

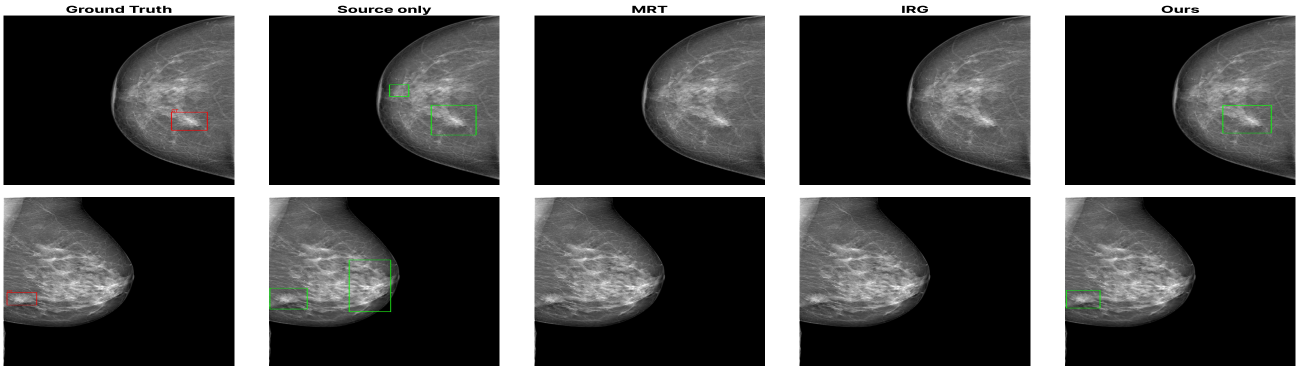

Figure 7. Quantitative results for Breast Cancer detection in `In2IB` setting. Predictions are depicted in green.