TITAN: Query-Token based Domain Adaptive Adversarial Learning

Tajamul Ashraf
MBZUAI
Masdar City, Abu Dhabi, UAE
tajamul.ashraf@mbzuai.ac.ae

Janibul Bashir
Gaash Lab, NIT Srinagar
Hazratbal, Jammu and Kashmir
janibbashir@nitsri.net

Abstract

*We focus on source-free domain adaptive object detection (SF-DAOD) problem, where the model has to adapt to an unlabelled target domain without using source data. Majority of existing frameworks for the problem employ a student-teacher (ST) framework where pseudo-labels are generated via a source-pretrained model for further fine-tuning. We observe that the performance of a student model often degrades drastically, due to the collapse of teacher model, primarily caused by high noise in pseudo-labels, resulting from domain bias, discrepancies, and a significant domain shift across domains. To obtain reliable pseudo-labels, we propose a Target-based Iterative Query-Token Adversarial Network (**TITAN**) which separates the target images into two subsets that are similar to the source (easy) and those that are dissimilar (hard). We propose a strategy to estimate variance to partition the target domain. This approach leverages the insight that higher detection variances correspond to higher recall and greater similarity to the source domain. Also, we incorporate query-token based adversarial modules into a student-teacher baseline framework to reduce the domain gaps between two feature representations. Experiments conducted on four natural imaging datasets and two challenging medical datasets have substantiated the superior performance of **TITAN** compared to existing state-of-the-art (SOTA) methodologies. We report an mAP improvement of +22.7, +22.2, +21.1, and +3.7 percent over the current SOTA on C2F, C2B, S2C, and K2C benchmarks, respectively. Code is available at https://github.com/Tajamul21/TITAN*

1. Introduction

Object Detection. Object detection is a well-studied problem in computer vision [6, 22, 101, 109, 118]. The advancement of deep learning methods in object detection has been significantly enabled by large-scale datasets with detailed annotations [15, 23, 27, 60, 105], which provide a solid foundation for supervised model training.

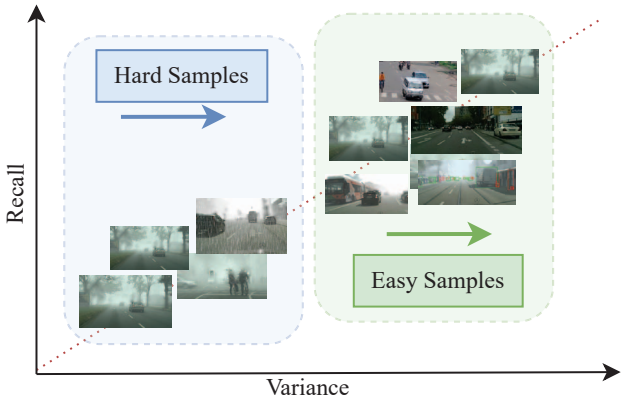

Figure 1. The core idea of our framework (**TITAN**) is that higher detection variances signal high recall and similarity to the source domain, enabling us to divide the target domain into easy and hard subsets.

Unsupervised Domain Adaptation (UDA). It has been widely observed that, despite their effectiveness in familiar visual contexts, deep object detection models often struggle to generalize to new visual domains. Unsupervised Domain Adaptation (UDA) has emerged as a widely adopted approach for addressing domain shifts [11, 25, 37, 38, 65, 79, 84], primarily by aligning the feature representations between the source and target domains [10, 36, 42, 43, 81, 87, 113]. Despite its effectiveness, a major drawback of this method is its reliance on access to source-domain data during the adaptation process, which can pose significant practical limitations [59, 63, 96]. To overcome this, our work concentrates on the more practical and challenging setting of Source-Free Domain Adaptive Object Detection (SF-DAOD).

Source-Free Domain Adaptive Object Detection (SF-DOAD). For image classification tasks, SF-DAOD has received significant attention in recent years [21, 40, 45, 95, 102, 103, 111]. However, there are relatively fewer works specifically addressing SF-DAOD [13, 40, 62, 71, 88]. Given the complexities of cluttered background, viewpoint variations, and many negative samples in an object detection

problem, directly applying traditional SFDA methods for classification tasks to SF-DAOD often leads to unsatisfactory performance.

Problems in Current SF-DOAD Approaches. Multiple state-of-the-art Source-Free Domain Adaptive Object Detection (SF-DAOD) methods adopt a self-supervised learning strategy based on a student-teacher (ST) paradigm. In these methods, pseudo-labels generated by a model trained on the source domain are used to guide the training of a student model [13, 53, 56, 86, 98]. However, when the source data contains inherent biases or there is a significant domain shift between the source and target domains, the quality of pseudo-labels deteriorates, leading to noisy supervision [19]. This noise can negatively affect the student model's learning. Furthermore, in such cases, the Exponential Moving Average (EMA) mechanism, which updates the teacher model using the student's weights, may propagate these errors, ultimately degrading the teacher model as well. Unlike in standard Unsupervised Domain Adaptation (UDA), where access to labeled source data helps stabilize training and mitigate pseudo-label noise, the SF-DAOD setting lacks this anchor, making it more susceptible to cumulative errors and model drift during adaptation.

Solution Strategies to Mitigate SF-DOAD Problems. To tackle the above issues in SF-DAOD recent techniques [13, 71] have proposed to use a larger update step size for EMA to slow down the teacher model's updating process deliberately. An alternative strategy involves emphasizing the past teacher model's influence by adjusting its contribution, thereby preserving previous knowledge and reducing the rate of model updates. However, such attempts have demonstrated limited effectiveness [62].

Our Insights and Proposed Strategy. To tackle this issue, we propose a query-token-driven adversarial learning approach (**TITAN**). Our method employs a variance-based detection strategy to separate target data into easy and challenging subsets, leveraging the insight that greater detection variance aligns with higher recall and stronger resemblance to the source domain, as illustrated in Fig. 1. Next, we integrate query-token-driven adversarial modules within a transformer-based student-teacher framework to bridge domain gaps in both local and instance-level feature representations, utilizing the FocalNet-DINO encoder and decoder accordingly.

Contributions. **(1)** We draw attention to a key challenge in the SF-DAOD setting, training instability of the student model arising from inaccurate pseudo-labels generated by the source-pretrained model. This issue becomes more severe when the source data contains inherent biases or when there exists a substantial domain gap between the source and target domains. **(2)** We identify a strong link between detection variance and resemblance to the source data. Leveraging this insight, we establish a method to categorize the target domain into easy and challenging subsets. **(3)** We introduce a query-token based adversarial alignment approach to refine the feature space, ensuring the generation of reliable pseudo-labels for the student-teacher framework. **(4)** We conduct extensive evaluation on four natural image adaptation benchmarks. We report an mAP of 50.2, 38.3, 59.8, and 53.2 on C2F, C2B, S2C, and K2C benchmarks respectively, against the performance of 40.9, 31.6, 49.4, and 51.3 by the current SOTA [13, 62]. **(5)** SF-DAOD is particularly relevant for medical imaging, where data sharing is restricted by privacy concerns and source data often suffers from single-center bias. We validate our approach on cross-domain mammogram datasets for Breast Cancer Detection (BCD), achieving state-of-the-art recall scores of 0.78 and 0.51 at 0.3 FPI on RSNA-BSD1K→INBreast and DDSM→INBreast, respectively, significantly outperforming the previous best (0.25 and 0.15) [87].

2. Related Work

Object Detection. In recent years, object detection, being a crucial computer vision task, has garnered considerable attention [12, 119]. Many object detection methods perform box regression and category classification using techniques such as anchors [61, 74, 75], proposals [28, 76], and points [83, 91, 115]. Traditional object detection systems heavily rely on extensive datasets like MS-COCO [60] and PASCALVOC [23], which necessitate significant time investment for annotation due to the large number of samples for each object category. Weakly supervised object detection (WSOD) methods [107, 108] leverage image-level labeled data [67] to train object detectors. In general, WSOD considers an image as a collection of region proposals and applies multiple instance learning to assign the image-level label to these proposals. [3, 4, 14, 94]. By using less expensive classification data, WSOD can expand the detection vocabulary without requiring costly instance-level annotations in object detection [17, 31, 31, 50, 89, 94, 116]. More recently, transformer-based models [1] have been developed for object detection, exploring token-wise dependencies for context modeling. Following the pioneering work of Vision Transformer (VIT) [22] and Detection Transformer (DETR) [6], transformers have emerged as a promising architecture in computer vision, demonstrating their efficacy in various tasks including object detection [6, 16, 101, 109, 114, 118]. Whereas most approaches emphasize supervised learning settings, our goal is to improve the model's ability to adapt to unseen domains without requiring extra annotations. To achieve this, we employ FocalNet-DINO (FND) [101] as our baseline detection model, chosen for its streamlined architecture and state-of-the-art transfer learning capability.

Source-Free Domain Adaptive Object Detection (SF-DOAD). In real-world applications, accessing source data during subsequent adaptation phases is often limited due to

privacy laws, data transfer restrictions, or proprietary concerns, especially in medical imaging. Source-Free Domain Adaptive Object Detection (SF-DAOD) enables knowledge transfer from pre-trained source models to the target domain without relying on sensitive source data [66]. The complexity of object detection, due to region diversity, multi-scale features, and deep architectures—combined with the lack of source data and reliable pseudo-labels, makes SF-DAOD significantly more challenging than traditional DAOD. As a result, SF-DAOD has emerged as a distinct and necessary focus within domain adaptation research. The Source-Free Object Detection (SFOD) method proposed in [56] leverages self-entropy descent to generate reliable pseudo-labels for self-training. SOAP [98] applies domain perturbations to the target data, enabling the model to learn domain-invariant features that remain robust against variations. LODS [53] incorporates a style enhancement module alongside a graph alignment constraint to promote the extraction of domain-agnostic features. A2SFOD [13] separates target samples into source-aligned and non-aligned groups and applies adversarial adaptation within a student-teacher setup. IRG [87] proposes an instance-level graph network with a contrastive learning objective to improve feature discrimination. PETS [62] enhances the ST framework by introducing a dynamic teacher model and a consensus module that fuses outputs from both static and dynamic teachers. However, these methods still struggle with training instability and local optima, largely due to relying on a single teacher update path.

3. Methodology

Source-free domain adaptive object detection (SF-DAOD) aims to adapt a detector, initially trained on a source domain, to an unlabeled target domain without direct access to source data. Given an unlabeled target dataset $\{X_i^t\}_{i=1}^N$ (where N denotes the total number of images) and a detector F initialized with source-trained parameters θ_s (e.g., a FocalNet-DINO model [101]), the objective is to refine these parameters to θ_t for effective performance in the target domain. In this work, we introduce a novel approach, **TITAN**, whose overall framework is depicted in Fig. 2. Our method categorizes target samples into two groups: source-similar (easy) and source-dissimilar (hard), using prediction variance from the source-trained detector F_{θ_s}. These groups are subsequently aligned using query-token-based adversarial learning within a ST framework. The following sections provide a detailed explanation of each stage.

3.1. Target Domain Division

Aligning the source and target domains is a central goal in domain adaptation [47, 92, 117]. This alignment can typically be achieved either in the input data space [8] or within the learned feature space [80]. However, in scenarios where source data is unavailable, domain alignment becomes considerably more challenging.

Although direct access to source data is restricted, the model trained on that domain inherently captures its underlying structure. To exploit this, we propose dividing the target dataset into two distinct subsets by evaluating how the pre-trained model responds to target instances. Specifically, we use a criterion based on prediction variance: the model's output variance on each target sample reflects how closely that sample resembles the source domain. A higher variance indicates greater uncertainty, which often occurs when the target sample shares characteristics with the source domain. Conversely, more familiar (and thus easier) target samples result in lower variance. This allows us to explicitly differentiate between source-similar and source-dissimilar target samples, using the pre-trained model as a proxy for source knowledge. The variance is calculated as:

$$v_i = \mathbb{E}[(F_{\theta_s}(X_i) - \mathbb{E}[F_{\theta_s}(X_i)])^2], \tag{1}$$

where, $F_{\theta_s}(X_i)$ denotes the prediction output for image X_i generated by the source-pretrained model. Computing the exact variance of these predictions is generally computationally intractable. To approximate this, we adopt a Monte Carlo sampling approach using dropout during inference, as introduced by Gal and Ghahramani [24]. Specifically, we perform M stochastic forward passes through the model while keeping its parameters fixed [5], allowing us to estimate the predictive uncertainty efficiently.

Since the outputs $F_{\theta_s}(X_i) = (\mathbf{b_i}, \mathbf{c_i})$ consist of bounding box coordinates and class scores, the detection variance is defined as the product of the bounding box variance v_{bi} and the class score variance v_{ci}. For a given prediction with N_i bounding boxes and K classes, where $\{\mathbf{b_{ij}} = (x_{ij}^1, y_{ij}^1, x_{ij}^2, y_{ij}^2)\}_{j=1}^{N_i}$ and $\{\mathbf{c_{ij}} = (c_{ij}^1, c_{ij}^2, \ldots, c_{ij}^K)\}_{j=1}^{N_i}$, we can express v_{bi} and v_{ci} as follows:

$$v_{bi} = \frac{1}{MN_i} \sum_{j=1}^{N_i} \sum_{m=1}^{M} \|\mathbf{b_{ij}^m} - \bar{\mathbf{b}}_{\mathbf{ij}}\|^2, \tag{2}$$

$$v_{ci} = \frac{1}{MN_i} \sum_{j=1}^{N_i} \sum_{m=1}^{M} \|\mathbf{c_{ij}^m} - \bar{\mathbf{c}}_{\mathbf{ij}}\|^2, \tag{3}$$

where $\mathbf{b_{ij}^m}$ and $\mathbf{c_{ij}^m}$ represent the localization coordinates and classification scores of the m-th forward pass for the j-th bounding box in X_i, respectively, and $\bar{\mathbf{b}}_{\mathbf{ij}}$, $\bar{\mathbf{c}}_{\mathbf{ij}}$ denote their corresponding average values over all M forward passes.

The detection variance for an image X_i is computed as $v_i = v_{bi} v_{ci}$. We then rank the images based on their variances, where r_i represents the rank of X_i. The variance level vl_i for the i-th image is given by $vl_i = \frac{r_i}{N}$. If $vl_i \geq \sigma$, the sample X_i is classified as source-similar; otherwise, it is treated as source-dissimilar, where $\sigma \in (0, 1)$ is a predefined threshold. This criterion enables us to split the target domain into two subsets—those resembling the source domain

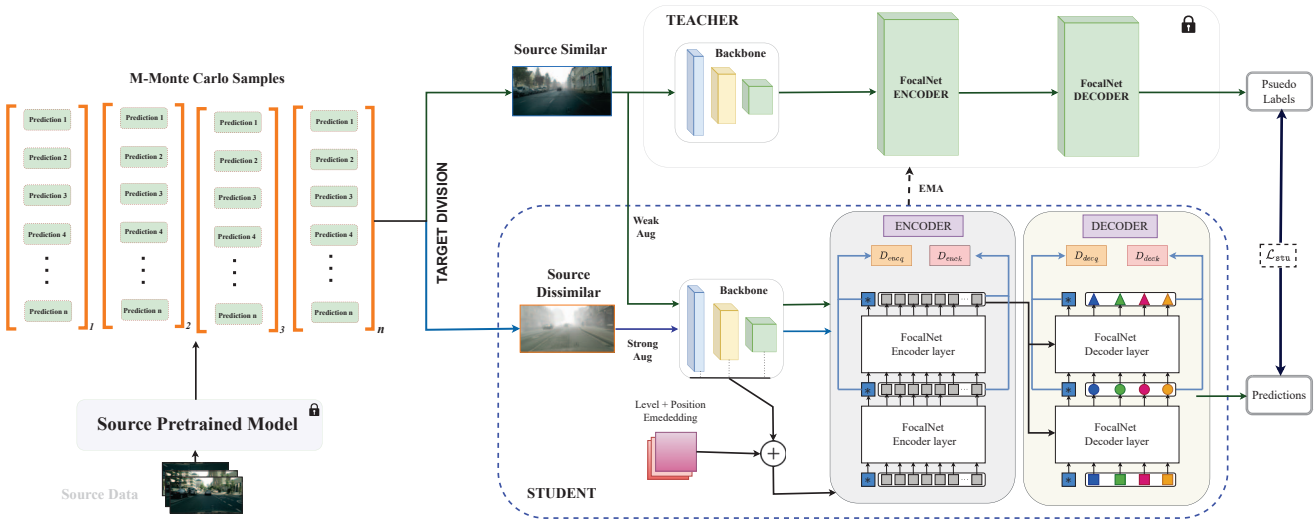

Figure 2. Overview: Target-based Iterative Query-Token Adversarial Network (**TITAN**)

and those that do not—facilitating query-driven adversarial alignment between domains.

3.2. Query-based Domain Adversarial Learning

We introduce query-based domain adversarial learning to align both easy and hard features globally. Specifically, on the encoder side, a query embedding q_d^{enc} is concatenated with the token sequence to form the input z_0 to the transformer encoder, *i.e.*,

$$z_0 = \left[q_d^{enc}; f_e^1; f_e^2; \cdots; f_e^N\right] + E_{pos} + E_{level}, \quad (4)$$

where $E_{pos} \in \mathbb{R}^{(N+1)\times C}$ is the positional embedding, $E_{level} \in \mathbb{R}^{(N+1)\times C}$ is the feature level embedding [118].

During the encoding phase, the query adaptively aggregates domain-specific features across the entire sequence. It captures the global context from the input images and focuses more on tokens with larger domain discrepancies. This query is then passed through a domain discriminator D_{enc_q} for efficient feature representations, *i.e.*,

$$\mathcal{L}_{enc_q}^{\ell} = d \log D_{enc_q}\left(z_\ell^0\right) + (1-d)\log\left(1 - D_{enc_q}\left(z_\ell^0\right)\right), \quad (5)$$

where $\ell = 1 \ldots L_{\text{enc}}$ denotes the layer indices in the encoder, and d is the domain label, taking the value 0 for source images and 1 for target images. Similarly, we append a query q_d^{dec} to the object queries to form the input sequence for the transformer decoder:

$$q_0 = \left[q_d^{dec}; q^1; q^2; \cdots; q^M\right] + E_{pos}^{'}, \quad (6)$$

where $E_{pos}^{'} \in \mathbb{R}^{(M+1)\times C}$ is the positional embedding and q^i is the i-th object query in the sequence. During the decoding phase, the query combines contextual information from each object query in the sequence, capturing the relationships among objects. This query is subsequently processed through the domain discriminator D_{dec_k} to obtain more effective feature representations.

$$\mathcal{L}_{dec_q}^{\ell} = d \log D_{dec_k}\left(q_\ell^0\right) + (1-d)\log\left(1 - D_{dec_q}\left(q_\ell^0\right)\right) \quad (7)$$

where $\ell = 1 \ldots L_{dec}$ indexes the layers in the transformer decoder.

> **The query-based domain adversarial learning facilitates better feature representations in both encoder and decoder but with distinct roles—aggregating global scene layouts in the encoder and encoding object relationships in the decoder. Leveraging attention and adversarial learning, it prioritizes aligning features with significant domain gaps while minimizing effort on well-aligned ones.**

3.3. Token-wise Domain Adversarial Learning

While query-based global adversarial learning effectively bridges the global domain gap in scene layout and inter-object relationships, it faces challenges in handling domain shifts due to local textures and styles. To address this, we introduce token-wise domain adversarial learning, which is applied to both the encoder and decoder of `FocalNet-DINO`.

In particular, each token embedding in the encoder sequence is passed through a domain classifier D_{enc_k} for adversarial training.

$$\mathcal{L}^{\ell}_{\text{enc}_k} = -\frac{1}{N}\sum_{i=1}^{N}\left[d\log D_{\text{enc}_k}(z^i_\ell) + (1-d)\log\left(1 - D_{\text{enc}_k}(z^i_\ell)\right)\right] \tag{8}$$

Likewise, a domain discriminator D_{dec_k} is applied on the decoder side to align each token embedding in the decoder, i.e.,

$$\mathcal{L}^{\ell}_{\text{dec}_k} = -\frac{1}{M}\sum_{i=1}^{M}\left[d\log D_{\text{dec}_k}(q^i_\ell) + (1-d)\log\left(1 - D_{\text{dec}_k}(q^i_\ell)\right)\right] \tag{9}$$

Query-based domain adversarial learning cannot be replaced by token-wise adversarial learning. While tokens in transformers focus more on local features due to their origin from image patches or object instances, domain queries aggregate global context without focusing on local details. This allows them to better address domain gaps related to scene layout and inter-object relationships.

While both the `FocalNet-DINO` encoder and decoder use token-wise adversarial learning, the encoder focuses on local texture and appearance, while the decoder aligns domain gaps at the object-instance level.

3.4. Cascaded Feature Alignment

To ensure thorough feature alignment, we implement cascaded feature alignment, progressively aligning source-similar and source-dissimilar features from shallow to deep layers. In the transformer encoder, this is expressed as:

$$\mathcal{L}_{enc} = \sum_{l=1}^{L_{enc}}\left(\mathcal{L}^{\ell}_{enc_k} + \lambda_{enc_q}\mathcal{L}^{\ell}_{enc_q}\right), \tag{10}$$

where λ_{enc_q} is a hyperparameter balancing query-based and token-based alignment losses, set to 0.1 in our experiments. Similarly, in the transformer decoder, we have:

$$\mathcal{L}_{dec} = \sum_{l=1}^{L_{dec}}\left(\mathcal{L}^{\ell}_{dec_k} + \lambda_{dec_q}\mathcal{L}^{\ell}_{dec_q}\right), \tag{11}$$

with λ_{dec_q} also set to 0.1. 3-layer MLPs are used as discriminators for both encoder and decoder, improving feature alignment.

3.5. Overall Objective

To summarize, the final training objective for **TITAN** is formulated as:

$$\min_G \max_D \mathcal{L}_{stu}(G) - \lambda_{enc}\mathcal{L}_{enc}(G,D) - \lambda_{dec}\mathcal{L}_{dec}(G,D) \tag{12}$$

Here, G represents the object detection model, and D refers to the domain discriminators. The hyperparameters λ_{enc}, λ_{dec}, and λ_{cons} control the relative importance of the different loss terms.

3.6. Generalization Analysis

Adversarial learning is applied to establish the transformation from the target domain to the source domain, and the success of this transformation is contingent on its generalization capabilities [32, 78].

Let μ denote the distribution of the original data, while ν represents the distribution of the generated data. The empirical approximations of these distributions are indicated as $\hat{\mu}_N$ and ν_N, where N is the size of the training sample. In adversarial training, the goal is to train a generator $g \in \mathcal{G}$ and a discriminator $f \in \mathcal{F}$, where $\mathcal{G}$ and $\mathcal{F}$ are the respective hypothesis spaces. The discriminator is implemented as a multilayer perceptron (MLP) with three layers, which consists of fully connected units and non-linear activation functions, as described in Section 3.4.

Theorem: *Covering Bound for the Discriminator*

Assume that the spectral norm of each weight matrix is bounded, i.e., $\|A_i\|_\sigma \leq s_i$, and that each weight matrix A_i has a corresponding reference matrix M_i such that $\|A_i - M_i\|_\sigma \leq b_i$ for $i = 1, \ldots, 3$. Then, the covering number $\mathcal{N}\left(\mathcal{F}|_S, \varepsilon, \|\cdot\|_2\right)$ satisfies the following inequality:

$$\log\mathcal{N}\left(\mathcal{F}|_S, \varepsilon, \|\cdot\|_2\right) \leq \frac{\log\left(2W^2\right)\|X\|_2^2}{\varepsilon^2}\prod_{i=1}^{3}s_i^2\sum_{i=1}^{3}\frac{b_i^2}{s_i^2}, \tag{13}$$

where W is the maximum dimension of the feature maps in the algorithm.

This result is derived from [2, 33], with a detailed proof provided in the Supplementary (§ A). The generalizability of GANs [30] depends on the complexity of the discriminator hypothesis [110]. Building on this, we utilize simple discriminators to enhance both generalizability and domain adaptation performance.

4. Experiments

4.1. Experimental Setup

Datasets.. The datasets used in our experiments include: **(1)** Cityscapes [15]: This dataset comprises urban scenes with 2,975 training images and 500 validation images. **(2)** Foggy Cityscapes [82]: Similar to Cityscapes, this dataset integrates fog and depth information into street view images. **(3)** KITTI [27]: A benchmark dataset for autonomous driving, containing images from real-world street scenes. For the experiments, only 7,481 training images were used. **(4)** SIM10k [46]: A synthetic dataset with 10,000 city scenery images of cars. **(5)** BDD100k [105]: An open-source video dataset for autonomous driving, including 100k images from various times, weather conditions, and driving scenarios. Its daytime subset consists of 36,728 training images and 5,258 validation images. **(6)** RSNA-BSD1K [7]: The original RSNA dataset [] includes 54,706 mammograms with 1,000 malignancies from 8,000 patients. RSNA-BSD1K is a

Table 1. : Results of adaptation from normal to foggy weather (C2F). "SF" refers to the source-free setting. “Oracle” refers to the models trained by using labels during training.

Method	Venue		Person	Rider	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
DA-Faster [9]	CVPR’18	✗	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
EPM [39]	ECCV’20	✗	44.2	46.6	58.5	24.8	45.2	29.1	28.6	34.6	39.0
SSAL [70]	NIPS’21	✗	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
SFA [90]	MM’21	✗	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
UMT [18]	CVPR’21	✗	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
D-adapt [44]	ICLR’21	✗	40.8	47.1	57.5	33.5	46.9	41.4	33.6	43.0	43.0
TIA [112]	CVPR’22	✗	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
PT [35]	ICML’22	✗	40.2	48.8	63.4	30.7	51.8	30.6	35.4	44.5	42.7
MTTrans [106]	ECCV’22	✗	47.7	49.9	65.2	25.8	45.9	33.8	32.6	46.5	43.4
SIGMA [54]	CVPR’22	✗	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
O2net [29]	MM’22	✗	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
AQT [42]	IJCAI’22	✗	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
AT [57]	CVPR’22	✗	43.7	54.1	62.3	31.9	54.4	49.3	35.2	47.9	47.4
TDD [34]	CVPR’22	✗	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	49.2
MRT [113]	ICCV’23	✗	52.8	51.7	68.7	35.9	58.1	**54.5**	41.0	47.1	51.2
HT [20]	CVPR’23	✗	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
CIGAR [64]	CVPR’23	✗	45.3	45.3	61.6	32.1	50.0	51.0	31.9	40.4	44.7
CSDA [26]	ICCV’23	✗	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8
SFOD [55]	AAAI’21	✓	21.7	44.0	40.4	32.6	11.8	25.3	34.5	34.3	30.6
SFOD-Mosaic [55]	AAAI’21	✓	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5
HCL [41]	NIPS’21	✓	26.9	46.0	41.3	33.0	25.0	28.1	35.9	40.7	34.6
SOAP [97]	IJIS’21	✓	35.9	45.0	48.4	23.9	37.2	24.3	31.8	37.9	35.5
LODS [53]	CVPR’22	✓	34.0	45.7	48.8	27.3	39.7	19.6	33.2	37.8	35.8
AASFOD [13]	AAAI’23	✓	32.3	44.1	44.6	28.1	34.3	29.0	31.8	38.9	35.4
IRG [87]	CVPR’23	✓	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
PETS [62]	ICCV’23	✓	46.1	52.8	63.4	21.8	46.7	25.5	37.4	48.4	40.3
LPLD [104]	ECCV’24	✓	39.7	49.1	56.6	29.6	46.3	26.4	36.1	43.6	40.9
TITAN (Ours)		✓	**52.8**	**51.7**	**68.0**	**43.2**	**65.5**	41.8	**46.0**	**48.7**	**52.2**
Oracle		✗	66.3	61.1	80.8	45.6	68.8	52.0	49.1	54.9	59.8

subset of 1,000 images with 200 malignant cases, annotated by expert radiologists. **(7)** INBreast [69]: A smaller BCD dataset with 410 mammography images from 115 patients, including 87 malignancies. **(8)** DDSM [51]: A publicly available BCD dataset comprising 2,620 full mammography images with 1162 malignancies. For full detail, please refer to the Supplementary material (§ B).

Task Settings. Building upon existing research [13, 40, 56, 62, 71, 113], we validate our method across four popular SF-DAOD and UDA benchmarks. In addition to this, we perform experiments on two cross-domain Medical Imaging (MI) tasks. These datasets represent various types of domain shifts, including[1]: **(1)** Cityscapes-to-Foggy-Cityscapes (C2F) **(2)** Cityscapes-to-BDD100k (C2B) **(3)** KITTI-to-Cityscapes (Car) (K2C) **(4)** Sim10k-to-Cityscapes (Car) (S2C) **(5)** RSNA-to-INBreast (R2In) **(6)** DDSM–to-INBreast (D2In)

Implementation Details. We adopt FocalNet-DINO (FND) [101] as our base detector. Initially, we set the loss coefficients to $\lambda_{\text{enc}} = 1.0$ and $\lambda_{dec} = 0.9$. The weight smoothing parameter α in the Exponential Moving Average (EMA) is configured to be 0.9996. The network is optimized using the Adam optimizer [49] with an initial learning rate of 2×10^{-4} and a batch size of 8. For data augmentation, we employ random horizontal flipping for basic augmentation and apply more advanced techniques such as random color jitter, grayscaling, and Gaussian blurring. The implementation is done using PyTorch [72]. The min-max loss function is realized through gradient reversal layers [25]. Further details are provided in Supplementary (§ C).

Evaluation Metrics. For natural image datasets, we report the average precision (AP) for each individual class along with the mean average precision (mAP) score, consistent with previous studies. For medical image datasets, we utilize the Free-Response Receiver Operating Characteristic (FROC) curves to evaluate detection performance, alongside F1-score and AUC for classification results. The FROC curves visually represent the trade-off between sensitivity/recall and false positives per image (FPI). We consider a prediction to be a true positive if the center of the predicted bounding box falls within the ground truth box [73].

4.2. Comparison with Current SOTA Methods

We evaluate the performance of our proposed **TITAN** method against other approaches on the four natural bench-

[1]The A-to-B (A2B) notation signifies the adaptation of a model pre-trained on the source domain A to the target domain B.

Table 2. Results of adaptation from small-scale to large-scale dataset (C2B).

Method	Venue		Person	Rider	Car	Truck	Bus	Mcycle	Bicycle	mAP
DA-Faster [9]	CVPR'18	✗	28.9	27.4	44.2	19.1	18.0	14.2	22.4	24.9
ICR [99]	CVPR'20	✗	32.8	29.3	45.8	22.7	20.6	14.9	25.5	27.4
EPM [39]	ECCV'20	✗	39.6	26.8	55.8	18.8	19.1	14.5	20.1	27.8
SFA [90]	MM'21	✗	40.2	27.6	57.5	19.1	23.4	15.4	19.2	28.9
AQT [42]	IJCAI'22	✗	38.2	33.0	58.4	17.3	18.4	16.9	23.5	29.4
ILLUME [48]	WACV'22	✗	33.2	20.5	47.8	20.8	33.8	24.4	26.7	29.6
O2net [29]	MM'22	✗	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
AWADA [68]	arXiv'22	✗	41.5	34.2	56.0	18.7	20.0	20.4	29.7	31.5
MTTrans [106]	ECCV'22	✗	44.1	30.1	61.5	25.1	26.9	17.7	23.0	32.6
MRT [113]	ICCV'23	✗	48.4	30.9	63.7	24.7	25.5	20.2	22.6	33.7
SFOD [55]	AAAI'21	✓	31.0	32.4	48.8	20.4	21.3	15.0	24.3	27.6
SFOD-M [55]	AAAI'21	✓	32.4	32.6	50.4	20.6	23.4	18.9	25.0	29.0
PETS [62]	ICCV'23	✓	42.6	34.5	62.4	19.3	16.9	17.0	26.3	31.3
AASFOD [13]	AAAI'23	✓	33.2	36.3	50.2	26.6	24.4	22.5	28.2	31.6
TITAN (Ours)		✓	**49.9**	**35.6**	**65.7**	**24.6**	**35.9**	**31.5**	29.2	**38.3**
Oracle			69.2	51.3	83.1	62.9	63.3	49.6	50.0	61.3

Table 3. (Left) Results of adaptation from synthetic to real scenes (S2C). (Right) Results of adaptation across cameras (K2C).

setting			Sim2City (S2C)	Kitty2City (K2C)
Method	**Venue**	**SF**	**Car (AP)**	**Car (AP)**
DA-Faster [9]	CVPR'18	✗	41.9	41.8
SAPNet [52]	ECCV'20	✗	44.9	43.4
EPM [39]	ECCV'20	✗	49.0	43.2
GPA [99]	CVPR'20	✗	47.6	47.9
MeGA-CDA [85]	CVPR'21	✗	44.8	43.0
DSS [93]	CVPR'21	✗	44.5	42.7
ViSGA [77]	ICCV'21	✗	49.3	47.6
SFA [90]	MM'21	✗	52.6	41.3
PT [35]	ICML'22	✗	-	55.1
MTTrans [106]	ECCV'22	✗	57.9	-
LODS [53]	CVPR'22	✗	-	43.9
CIGAR [64]	CVPR'23	✗	58.5	48.5
CSDA [26]	ICCV'23	✗	57.8	48.6
SFOD [55]	AAAI'21	✓	42.3	43.6
SFOD-Mosaic [55]	AAAI'21	✓	42.9	44.6
IRG [87]	CVPR'21	✓	45.2	46.9
AASFOD [13]	AAAI'23	✓	44.0	44.9
PETS [62]	ICCV'23	✓	57.8	47.0
LPLD [104]	ECCV'24	✓	49.4	51.3
(Ours)		✓	**59.8**	**53.2**
Oracle		✗	63.9	62.1

marks and two medical benchmarks mentioned earlier. Since UDA and SF-DAOD share similar task settings, we conducted comparisons with both. Table 1-3 and Table 4 present the comparison results on natural and medical images, respectively. Our proposed **TITAN** consistently outperforms existing state-of-the-art (SOTA) methods, demonstrating significant improvements across both natural and medical images, highlighting its effectiveness in both domains and its potential to advance performance in real-world applications. This further establishes **TITAN** as a promising solution for tackling domain adaptation challenges in various tasks.

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
R2In	D-adapt [44]	ICLR'21	✗	0.04	0.12	0.18	0.29	0.439	0.263
	AT [58]	CVPR'22	✗	0.16	0.28	0.35	0.42	0.486	0.338
	H2FA [100]	CVPR'22	✗	0.03	0.13	0.18	0.36	0.634	0.236
	MRT [113]	ICCV'23	✗	0.32	0.52	0.69	0.72	0.741	0.352
	Mexformer [90]	MM'21	✓	0.24	0.31	0.39	0.39	0.336	0.287
	IRG [19]	CVPR'23	✓	0.16	0.25	0.37	0.39	0.308	0.235
	LPLD [104]	ECCV'24	✓	0.25	0.25	0.45	0.43	0.548	0.635
	Ours			**0.59**	**0.78**	**0.80**	**0.83**	**0.892**	**0.850**

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
D2In	D-adapt [44]	ICLR'21	✗	0.00	0.06	0.09	0.1	0.381	0.362
	AT [58]	CVPR'22	✗	0.01	0.08	0.10	0.15	0.385	0.311
	H2FA [100]	CVPR'22	✗	0.02	0.08	0.10	0.12	0.483	0.315
	MRT [113]	ICCV'23	✗	0.03	0.09	0.12	0.17	0.739	0.587
	Mexformer [90]	MM'21	✓	0.02	0.03	0.03	0.03	0.06	0.09
	IRG [19]	CVPR'23	✓	0.05	0.05	0.07	0.09	0.11	0.12
	LPLD [104]	ECCV'24	✓	0.09	0.15	0.35	0.35	0.548	0.635
	Ours			**0.36**	**0.51**	**0.75**	**0.81**	**0.825**	**0.838**

Table 4. (left-top) Results on adaptation from large to small-scale medical datasets with different modalities (R2In), (left-bottom) Results on adaptation across medical datasets with different machine-acquisitions (D2In)

4.2.1. Adaptation on Natural Images

1. (C2F): Adaptation results from clear to foggy weather shown in Tab. 1. Our method surpasses both UDA and SF-DAOD methods.
2. (C2B): Adaptation from small to large-scale datasets. Please see Tab. 2.
3. (K2C): Adaptation across different cameras. Tab. 3 evaluates our model's performance on domain shifts due to different camera settings viz. resolution, quality etc.
4. (S2C): Adaptation from synthetic to real images. To explore adaptation from synthetic to real scenes, we employ a model pretrained on the complete **SIM10k** dataset [46] as the source model. We then adapt this model to the **Cityscapes** dataset [15], with only car images retained and other categories discarded as shown in Tab. 3. Further details are provided in the Supplementary (§ B).

(a) domain query-based feature alignment **on the encoder side** **(b)** domain query-based feature alignment **on the decoder side**

Figure 3. "Visualization of domain queries from both the encoder and decoder sides of FocalNet-DINO, in the C2F scenario."

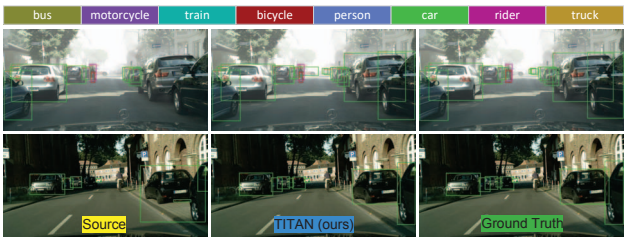

Figure 4. Qualitative results comparing Source model (FocalNet-DINO), **TITAN** (ours), and ground truth visualizations. Our method shows accurate predictions.

4.2.2. Adaptation on Medical Imaging Datasets

1. (R2In): Adapting across datasets with different machine acquisitions, using RSNA-BSD1K [7] as source and INBreast [69] as the target. Our method outperforms existing approaches across FPI values (Tab. 4).
2. (D2In): Adapting from large to small-scale medical datasets with different modalities, from **DDSM** [51] to **INBreast** [69]. Our method achieves SOTA results across FPI values (Tab. 4).

4.3. Ablation Studies

To gain deeper insights into our method, we perform ablation studies by isolating each component of **TITAN**, as presented in Table 5. We have the following observations: (1) both domain query-based adversarial learning and token-wise adversarial learning can alleviate the domain gaps and improve FocalNet-DINO transformer's cross-domain performance by 40.8 and 39.7 mAP, respectively; (2) query-based domain adversarial learning and token-wise domain adversarial learning are complementary to each other. Thereby, a combination of both brings further improvement; (3) The target division (TD) based adversarial learning is effective, resulting in a gain of 51.2 mAP.

4.3.1. Visualization of Domain Query.

The query-token domain feature alignment adaptively aggregates global context in the encoder and decoder. As visualized in Fig. 3, the encoder query attends to regions with domain gaps, like dense fog, while the decoder query emphasizes foreground objects such as cars and bicycles, capturing key domain shifts. More visualizations and ablations are provided in the Supplementary Material (§ F).

Table 5. Ablation studies on query-token based domain adversarial learning, **without target division**, are conducted on the C2F scenario. DQ_{enc} and DQ_{dec} are applied to the final encoder and decoder layers, respectively. Similarly, TW_{enc} and TW_{dec} denote **TITAN** applied to the last encoder and decoder layers.

Baseline	DQ_{enc}	DQ_{dec}	TW_{enc}	TW_{dec}	TD	mAP
✓						35.2
	✓					37.5
		✓				36.7
			✓			38.5
				✓		37.9
	✓	✓				40.8
			✓	✓		39.7
	✓		✓			47.1
		✓		✓		45.6
✓	✓	✓	✓	✓	✓	**51.2**

4.3.2. Detection Results

Fig. 4 presents in-depth visual comparisons between FocalNet-DINO and our framework **TITAN**, alongside the ground-truth annotations. As shown, **TITAN** consistently enhances detection performance across all three scenarios. It effectively reduces false positives and identifies difficult objects that FocalNet-DINO misses. In the first row, **TITAN** successfully detects a car that is distant and not labeled in the ground-truth, highlighting its strong generalization capability to the target domain.

5. Conclusion

We introduce a novel source-free domain adaptive object detection framework that segments the target domain into easy and hard samples, and subsequently aligns them in the feature space through query-token-based adversarial learning. Extensive experiments on both natural and medical image benchmarks show that our method consistently surpasses existing state-of-the-art approaches. Additionally, ablation studies validate the contribution of each component to enhancing the model's ability to adapt across domains.

References

[1] Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017. 2

[2] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017. 5, 1

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2846–2854, 2016. 2

[4] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1081–1089, 2015. 2

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, page 1613–1622, 2015. 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2

[7] C Carr, F Kitamura, J K-Cramer, J Mongan, K Andriole, M V, M Riopel, R Ball, and S Dane. Rsna screening mammography breast cancer detection. 2022. 5, 8, 3

[8] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8866–8875, 2020. 3

[9] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[10] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1

[11] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 1

[12] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[13] Qiaosong Chu, Shuyan Li, Guangyi Chen, Kai Li, and Xiu Li. Adversarial alignment for source free object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 452–460, 2023. 1, 2, 3, 6, 7

[14] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016. 2

[15] 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, 5, 7

[16] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[18] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, 2021. 6

[19] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2, 7

[20] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23829–23838, 2023. 6

[21] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. Confident anchor-induced multi-source free domain adaptation. *Advances in Neural Information Processing Systems*, 34:2848–2860, 2021. 1

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[23] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 2

[24] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 3

[25] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1, 6

[26] Changlong Gao, Chengxu Liu, Yujie Dun, and Xueming Qian. Csda: Learning category-scale joint feature for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11421–11430, 2023. 6, 7

[27] 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, 5, 2

[28] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[29] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. Improving transferability for domain adaptive detection transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1543–1551, 2022. 6, 7

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[31] Cagri Gungor and Adriana Kovashka. Boosting weakly supervised object detection using fusion and priors from hallucinated depth. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2024. 2

[32] Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020. 5

[33] Fengxiang He, Tongliang Liu, and Dacheng Tao. Why resnet works? residuals generalize. *IEEE transactions on neural networks and learning systems*, 31(12):5349–5362, 2020. 5

[34] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9570–9580, 2022. 6

[35] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9570–9580, 2022. 6, 7

[36] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6668–6677, 2019. 1

[37] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 1

[38] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018. 1

[39] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 733–748. Springer, 2020. 6, 7

[40] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 1, 6

[41] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data, 2022. 6

[42] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *IJCAI*, pages 972–979, 2022. 1, 6, 7

[43] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1

[44] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection, 2022. 6, 7

[45] Mengmeng Jing, Xiantong Zhen, Jingjing Li, and Cees Snoek. Variational model perturbation for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:17173–17187, 2022. 1

[46] 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. 5, 7, 2

[47] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019. 3

[48] Vaishnavi Khindkar, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, Rohit Saluja, and CV Jawahar. To miss-attend is to misalign! residual self-attentive feature alignment for adapting object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3632–3642, 2022. 7

[49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[50] Hannah Kniesel, Leon Sick, Tristan Payer, Tim Bergner, Kavitha Shaga Devan, Clarissa Read, Paul Walther, Timo Ropinski, and Pedro Hermosilla. Weakly supervised virus capsid detection with image-level annotations in electron microscopy images. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[51] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017. 6, 8, 3

[52] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation, 2020. 7

[53] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2022. 2, 3, 6, 7

[54] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5291–5300, 2022. 6

[55] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data, 2020. 6, 7

[56] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8474–8481, 2021. 2, 3, 6

[57] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 6

[58] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection, 2022. 7

[59] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 1

[60] 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. 1, 2

[61] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2

[62] Qipeng Liu, Luojun Lin, Zhifeng Shen, and Zhifeng Yang. Periodically exchange teacher-student for source-free object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6414–6424, 2023. 1, 2, 3, 6, 7

[63] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 1

[64] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23776–23786, 2023. 6, 7

[65] Shao-Yuan Lo, Wei Wang, Jim Thomas, Jingjing Zheng, Vishal M Patel, and Cheng-Hao Kuo. Learning feature decomposition for domain adaptive monocular depth estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8376–8382. IEEE, 2022. 1

[66] Xin Luo, Wei Chen, Zhengfa Liang, Longqi Yang, Siwei Wang, and Chen Li. Crots: Cross-domain teacher–student learning for source-free domain adaptive semantic segmentation. *International Journal of Computer Vision*, 132(1): 20–39, 2024. 3

[67] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[68] Maximilian Menke, Thomas Wenzel, and Andreas Schwung. Awada: Attention-weighted adversarial domain adaptation for object detection, 2022. 7

[69] 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. 6, 8, 2

[70] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *Advances in Neural Information Processing Systems*, 34:22770–22782, 2021. 6

[71] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 6

[72] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[73] Krithika Rangarajan, Pranjal Aggarwal, Dhruv Kumar Gupta, Rohan Dhanakshirur, Akhil Baby, Chandan Pal, Arun Kumar Gupta, Smriti Hari, Subhashis Banerjee, and Chetan Arora. Deep learning for detection of iso-dense, obscure masses in mammographically dense breasts. *European radiology*, 2023. 6

[74] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[75] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[76] 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

[77] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9204–9213, 2021. 7

[78] Afshin Rostamizadeh, Ameet Talwalkar, and Mehryar Mohri. Foundations of machine learning, 2012. 5

[79] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1

[80] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 3

[81] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6956–6965, 2019. 1

[82] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5, 1, 2

[83] Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in Neural Information Processing Systems*, 35:34899–34911, 2022. 2

[84] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

[85] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection, 2021. 7

[86] Vibashan Vs, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mixture of teacher experts for source-free domain adaptive object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3606–3610. IEEE, 2022. 2

[87] Vibashan VS, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3520–3530, 2023. 1, 2, 3, 6, 7

[88] Vibashan VS, Poojan Oza, and Vishal M Patel. Towards online domain adaptive object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 478–488, 2023. 1

[89] Jian Wang, Liang Qiao, Shichong Zhou, Jin Zhou, Jun Wang, Juncheng Li, Shihui Ying, Cai Chang, and Jun Shi. Weakly supervised lesion detection and diagnosis for breast cancers with partially annotated ultrasound images. *IEEE Transactions on Medical Imaging*, 2024. 2

[90] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021. 6, 7

[91] Xinggang Wang, Kaibing Chen, Zilong Huang, Cong Yao, and Wenyu Liu. Point linking network for object detection. *arXiv preprint arXiv:1706.03646*, 2017. 2

[92] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *CVPR*, pages 7289–7298, 2019. 3

[93] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection, 2021. 7

[94] Yuting Wang, Velibor Ilic, Jiatong Li, Branislav Kisačanin, and Vladimir Pavlovic. Alwod: Active learning for weakly-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6459–6469, 2023. 2

[95] Zhenbin Wang, Mao Ye, Xiatian Zhu, Liuhan Peng, Liang Tian, and Yingying Zhu. Metateacher: Coordinating multi-model domain adaptation for medical image classification. *Advances in Neural Information Processing Systems*, 35: 20823–20837, 2022. 1

[96] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9010–9019, 2021. 1

[97] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems*, 36, 2021. 6

[98] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems*, 36(8):3746–3766, 2021. 2, 3

[99] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection, 2020. 7

[100] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14309–14319, 2022. 7

[101] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 1, 2, 3, 6

[102] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021. 1

[103] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022. 1

[104] Ilhoon Yoon, Hyeongjun Kwon, Jin Kim, Junyoung Park, Hyunsung Jang, and Kwanghoon Sohn. Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In *European Conference on Computer Vision*, pages 337–353. Springer, 2024. 6, 7

[105] 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, 5, 2

[106] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, pages 629–645. Springer, 2022. 6, 7

[107] Dingwen Zhang, Wenyuan Zeng, Jieru Yao, and Junwei Han. Weakly supervised object detection using proposal-and semantic-level relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3349–3363, 2020. 2

[108] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021. 2

[109] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

[110] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017. 5

[111] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. *Advances in Neural Information Processing Systems*, 35: 5137–5149, 2022. 1

[112] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14217–14226, 2022. 6

[113] Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, and Yang Liu. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19039–19049, 2023. 1, 6, 7

[114] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 2

[115] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[116] 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*, pages 350–368. Springer, 2022. 2

[117] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 3

[118] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 4

[119] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 2