Context Aware Grounded Teacher for Source Free Object Detection

Tajamul Ashraf[1,2*], Rajes Manna[4†], Partha Sarathi Purkayastha[3†], Tavaheed Tariq[4], Janibul Bashir[4*]

[1*]Department of Computer Vision, MBZUAI, Masdar City, 11058, Abu Dhabi, UAE.
[2*]School of Information Technology, IIT Delhi, Hauz Khas, 180037, New Delhi, India.
[3*]Microsoft Research India, Bengaluru, 560001, Karnataka, India.
[4*]Department of Information Technology, NIT Srinagar, Hazratbal, 190007, J&K, India.

*Corresponding author(s). E-mail(s): tajamul.ashraf@mbzuai.ac.ae; janibbashir@nitsri.ac.in;
Contributing authors: rajes_2021bite063@nitsri.ac.in; t-ppurkayast@microsoft.com; tavaheed_2022bite008@nitsri.ac.in;
†These authors contributed equally to this work.

Abstract

Source-free domain adaptation for object detection (SFOD) remains a challenging paradigm due to the dual impact of context bias arising from class imbalance and inter-class correlations, and the instability of teacher-student frameworks under noisy pseudo-labels associated with domain shifts. To tackle the problem of context bias and the significant performance drop of the student model in the `SFOD` setting, we introduce Grounded Teacher (`GT`), a bias-aware semi-supervised adaptation framework that grounds the teacher model through relational and semantic regularization, as a standard framework. GT introduces a Relational Context Module (RCM) to explicitly model directional confusions between classes, maintaining an exponential moving average (EMA) estimate of cross-domain contextual bias. Building upon this, a Semantic Augmentation (SA) strategy selectively augments minority and confusable classes through adaptive MixUp in both source-similar and source-dissimilar target regions, improving minority recall without overfitting dominant categories. To stabilize learning under biased pseudo-labels, we design a Semantic-Aware Loss (SAL) that applies diagonally normalized weights, preventing gradient explosion while emphasizing minority–majority corrections. Additionally, a frozen Expert branch derived from large vision foundation models (LVFMs) serves as a supervisory reference during training, refining pseudo-label quality without adding inference overhead. GT's behavior-driven bias quantification makes it broadly applicable across domains without relying on dataset priors. Evaluations on standard SFOD benchmarks of Cityscapes→Foggy and medical transfers (DDSMINBreast, RSNA) demonstrate consistent performance gains, achieving significant improvements in minority-class detection. Extensive ablations further validate the contribution of each component and the computational efficiency of GT. All relevant resources, including preprocessed data, trained model weights, and code, are publicly available at this link.

Keywords: Generalization, Source-Free Domain Adaptation, Context Bias, Semi-Supervised Learning, Large Vision Foundation Models, Object Detection

1 Introduction

Unsupervised Domain Adaptation (UDA) offers a strategy for adapting object detectors to new domains where labeled data is not available.

Moreover many popular SFOD techniques utilize a self-supervised approach in a student-teacher (ST) framework. These methods bootstrap by training on pseudo-labels generated by a source-pretrained model [67, 123, 70, 21, 112]. However, if the source data is biased, or the domain shift between source and target domains is significant, there is noise in the pseudo-labels, which impacts the training of a student model [27]. Since the pseudo-label error is significant, the, Exponential Moving Average (EMA) step, which updates the teacher model from the student model's weight, ends up corrupting the teacher model as well leading to a mode collapse [3]. This phenomenon of progressive degradation, especially in source-free setting where source data is not available during adaptation, is visually evident in Figure 1, where traditional Mean Teacher frameworks and SF-Student-Teacher frameworks suffer from declining detection quality over time, while our method remains robust over the epochs.

Prior works have documented pseudo-label noise and EMA instabilities in teacher–student adaptation, including in SFOD baselines and variants inspired by Unbiased Teacher/Mean-Teacher families. Our contribution is not simply the observation of these issues, but a coherent, source-free solution to address them. We introduce our Grounded Teacher (GT), specifically designed to tackle context imbalance and mode collapse in the SFOD setting, which comprises a behavioral confusion estimator (RCM) with dual use in data-space augmentation (SA) and loss-space reweighting (SAL), along with training-only LVFM Expert guidance to stabilize the teacher. Together, these mechanisms target context bias and teacher drift without any source images or inference-time cost.

The Relation-Contextual Module (RCM) explicitly models class biases and inter-class dynamics within the training data. Leveraging prior insights, we implement a semantic augmentation strategy that enhances minority class representation by strategically blending them with visually similar

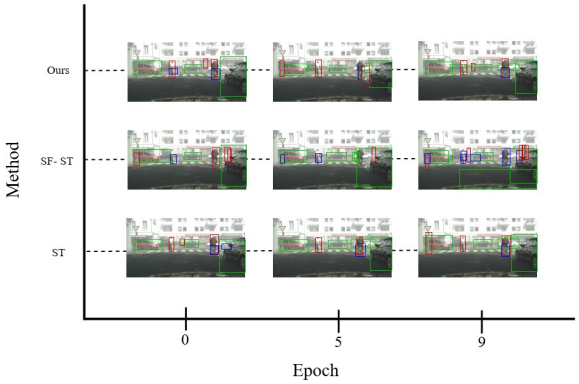

Fig. 1: *Illustrative effect of pseudo-label noise and EMA drift in source-free object detection. Each row shows representative target-domain detections (Cityscapes→Foggy) across training epochs for (a) Mean Teacher (ST), (b) its Source-Free variant (SF-ST), and (c) our Grounded Teacher (GT). Conventional mean-teacher frameworks experience progressive drift as pseudo-labels introduce false positives and obscure minority classes, causing instability. GT mitigates this through its RCM, SA/SAL and Expert Supervision, which jointly manage class-confusion dynamics, ensuring consistent, spatially accurate detections and stable performance across epochs under source-free conditions.*

majority class instances, drawn from our Cropbank repository[140]. This augmentation operates bidirectionally across domains, simultaneously addressing both class imbalance and domain shift. To further mitigate inter-class bias, we propose a Semantic-Aware Loss that dynamically prioritizes minority classes during training, particularly focusing on challenging cases where they are most vulnerable to misclassification as majority classes. Together, these components form a comprehensive framework that holistically addresses representation learning challenges in cross-domain scenarios.

Another key insight of this paper is that one can use learnt representations from Vision Foundational Models to guide a student model in SFOD settings. Due to the excellent zero shot capabilities of the Vision Foundation Models (VFMs), they can effectively guide student models to learn the right representation, and distances between the samples, even when the target domain has a large shift from the source domain. This helps the model overcome mode collapse caused by biased pseudo-labels during unsupervised training.

Through the integration of these approaches, we observe enhanced pseudo-label quality, leading to measurable gains on medical imaging benchmarks achieving a new state-of-the-art (SOTA) of 50.8 mAP in natural images for Cityscapes → Foggy Cityscapes, beating the previous standard of 45.0 mAP by +5.8 mAP (∼13% performance gain). We summarize the contributions of this paper as follows:

- We highlight the problem of training instability of a student model in SFOD setting due to incorrect pseudo labels from a source pretrained model in the presence of source bias, or large domain shift between source and target domains.
- We propose our GT model, supported by our RCM, which can map the model's existing class biases. To this end, we formulate a Semantic Aware Loss (SAL) to further enhance the performance of biased classes.
- We propose to leverage the zero-shot performance of VFMs to guide the student model in a sample-specific manner and help learn the correct representation in a large domain shift scenario.

2 Related Works

Source-free object detection (SFOD) adapts a detector to a target domain without accessing source images. In practice, most methods build upon the mean-teacher family which involves a student learning from pseudo-labels while a teacher is updated by EMA, where accuracy rises or drifts depending on pseudo-label quality and class imbalance. Our work follows this line but adds two design principles: (i) measure the directed minority→majority confusions as they actually occur during adaptation, and (ii) act on those measurements once in the data space (semantic augmentation) and once in the loss space (bounded reweighting), while allowing training-only guidance from large vision foundation models (LVFMs).

2.1 Relation-aware teachers and class-bias remedies (DAOD vs. SFOD)

A close line of work in domain-adaptive detection (DAOD) argues that inter-class relations can be exploited to counter class bias when both source and target images are available. CAT (Class-Aware Teacher) argues that inter-class relations can be exploited to counter class bias when both source and target images are available. CAT learns an Inter-Class Relation module, uses those relations to drive class-relation augmentation, and introduces a class-relation loss. The approach is trained with source data and reported on Cityscapes→Foggy and related shifts, where the results support the premise that explicitly modeling relations helps close minority-class gaps, but its mechanisms assume access to source supervision during adaptation.

By contrast, IRG treats source-free detection and builds an Instance Relation Graph on target proposals to guide a contrastive objective inside a mean-teacher loop. IRG's graph therefore is the supervisory signal that shapes representations, without performing confusion-conditioned augmentation or class-wise bounded reweighting. Our method remains in the source-free regime like IRG, but it differs in what is modeled (a behavioral, directed confusion matrix rather than a latent graph), how it is used (to steer MixUp pairing and loss weights), and where cost is paid (LVFM guidance only at training time, no inference overhead).

2.2 Contemporary SFOD under mean-teacher dynamics

Recent SFOD work simplifies or augments the idea of the mean-teacher framework. SF-UT (Source-Free Unbiased Teacher) streamlines teacher–student training with strong/weak augmentation and careful pseudo-label selection, reporting robust baselines across urban benchmarks. This highlights that much of SFOD's gain still depends on which pseudo-labels are to be trusted. LPLD (Low-confidence Pseudo-Label Distillation) goes further by distilling low-confidence pseudo-labels mined from RPN proposals, explicitly rescuing small or hard instances that high-confidence thresholds miss. While doing

so, it also uses class-relation cues when filtering low-confidence candidates to help reduce noise and improve performance. Orthogonal to both, Dynamic Retraining-Updating (DRU) revisits its teacher–student coordination by adaptively retraining and updating the network to mitigate drift. It monitors student uncertainty through decoder variance as it selectively reinitializes stagnating decoder layers and dynamically tunes the teacher's EMA update rate, while using historical loss regularization to stabilize learning. These works point to three distinct aspects of focus: pseudo-label selection, coverage of hard instances, and teacher update policy. Grounded Teacher complements them with a fourth aspect: confusion-conditioned data and loss interventions along with training-only expert guidance, a combination aimed specifically at minority→majority drift while keeping deployment cost unchanged.

2.3 Adjacent work in segmentation

Outside detection, DRSL (Distribution Regularized Self-Supervised Learning) regularizes pixel-level multi-modal class distributions for domain-adaptive segmentation and aligns source+target modes to reduce pseudo-label noise. DRSL is informative because it too addresses distributional noise, but operates with a different objective to minimze intra-class distribution mismatch between source and target domains at the pixel level by aligning multi-modal latent features for each class, while preserving class consistency across domains. In contrast, GT's objective operates instance-wise and source-free, targeting inter-class confusion reduction rather than cross-domain distribution alignment. Our detector estimates class-wise confusion from target pseudo-labels, route augmentation through variance-split Cropbanks, and apply a diagonal-normalized, bounded loss that emphasizes minority→majority confusions without gradient blow-up. The two perspectives are thus complementary rather than interchangeable across tasks.

2.4 Rationale for GT

GT draws on lessons from all the prior works addressed above but differs in execution. We measure bias as a directed confusion process, use it twice (augmentation and bounded reweighting), and utilize expertise from LVFMs only while training, so the deployed detector remains unchanged. This places GT as a source-free, instance-level, confusion-guided alternative that complements pseudo-label selection and teacher-update schedules, and empirically strengthens minority-class consistency without inference-time cost. (Sections §3–§4 develop these choices and quantify their effects.)

Our setting inherits the Mean Teacher idea that weight-averaged teachers provide stable targets under consistency training, and the Unbiased Teacher insight that pseudo-label bias must be controlled in detection. GT stays within this paradigm but focuses primarily on measuring and regulating confusions during source-free adaptation, with LVFMs as expert supervisors rather than as deployed detectors.

3 Preliminaries

Problem statement. Let X denotes the input space and $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$, the source domain, where $y_s^i = (\mathbf{b_s^i}, c_s^i)$ is a tuple comprising of ground truth bounding box ($\mathbf{b_s^i} \in \mathcal{R}^4$) and the corresponding class label ($c_s^i \in \mathcal{R}$) of the object respectively. Here, $x_s \in X$ denotes the image from the source domain, and N_s denotes the total number of source images. The target domain $\mathcal{D}_t = \{x_t^i\}_{i=1}^{N_t}$ is unlabeled, where N_t denotes the total number of target images, and the target samples $x_t^i \in X$. Our objective is to transfer the source-trained model to the target domain without accessing samples from the source domain.

Mean Teacher Framework. The self-supervised adaptation strategy involves updating the student model using unlabeled target data, leveraging pseudo-labels generated from the teacher model. Both the student and the teacher models are initialized by the source-trained model. The pseudo labels undergo a confidence threshold-based filtering process, and only the reliable ones are utilized to supervise the student training [62]. The pseudo-label supervision loss for the object detection model can be expressed as:

$$\mathcal{L}_{\text{stu}} = \mathcal{L}^S_{\text{box}}(x_t^i, yt_t^i) + \mathcal{L}^S_{\text{giou}}(x_t^i, yt_t^i) + \mathcal{L}^S_{\text{cls}}(x_t^i, yt_t^i). \quad (1)$$

Here yt_t^i represents the pseudo-label for image i obtained after filtering low-confidence predictions

from the teacher. The teacher is updated progressively through an Exponential Moving Average (EMA) of student weights. We implement a hard threshold, τ, on the classification scores generated by the teacher to ensure that only pseudo-labels with high confidence are utilized by the student network, thereby encouraging more reliable learning outcomes.

Thus, the overall self-training process for the traditional student-teacher (ST) based object detection framework is expressed as follows:

$$\Theta_s \leftarrow \Theta_s + \gamma \frac{\partial L_{\text{stu}}}{\partial \Theta_s}, \tag{2}$$

$$\Theta_t \leftarrow \alpha \Theta_t + (1 - \alpha)\Theta_s. \tag{3}$$

Here $\mathcal{L}_{\text{stu}}$ represents the student loss computed using pseudo-labels from the teacher network, and Θ_s and Θ_t denote student and teacher network, respectively. The parameters γ and α denote the student learning rate and teacher EMA rate, respectively. Although the ST framework enables knowledge distillation using noisy pseudo-labels, it alone is insufficient for learning high-quality target features in a source-free setting and leads to model degradation. Therefore, to enhance the features in the target domain, we introduce Visual Foundation Models (VFMs) as Expert and incorporate supervised loss for knowledge transfer from expert VFMs. Following [72], a discriminator is added to encourage domain invariant feature representations with an associated loss, $\mathcal{L}_{\text{dis}}$.

4 Methodology

Our proposed method, Grounded Teacher (GT), illustrated in Figure 3, extends the standard mean-teacher framework by introducing a novel Relation Contextual Module (RCM). At the heart of GT, the RCM is designed to systematically capture and quantify class-specific biases in the model. Unlike conventional methods that address class imbalance in a general way, RCM explicitly models the semantic relationships between classes, with a particular focus on minority classes that are frequently misclassified as dominant ones.

This is achieved through the construction of a batch-wise confusion matrix based on pseudo-labels, which is normalized and continuously aggregated into a global matrix. This evolving matrix provides a robust and dynamic representation of the model's bias landscape, forming the foundation for two key components of our method: Class-Relation Augmentation and Semantic Loss. Class-Relation Augmentation targets image-level class imbalance. Using the class relationships encoded by RCM, we identify minority classes in an image that share strong semantic overlap with dominant classes. To increase their representation and improve discrimination, we apply MixUp [142]—a technique known to be effective in imbalanced scenarios [36, 20]—to blend instances of these semantically related classes. This not only amplifies the presence of minority class samples but also encourages the model to learn finer distinctions between closely related categories (Figure 3).

In parallel, the RCM guides the formulation of our Semantic-Aware Loss, a weighted loss function that assigns greater importance to minority classes—especially those prone to misclassification. By amplifying the penalty for errors on these classes, our approach effectively mitigates the model's inherent biases and promotes balanced, equitable learning across all categories.

4.1 Relation Contextual module

Previous efforts in addressing class imbalance within domain adaptation settings [50, 56, 104] have contributed significantly to improving performance on underrepresented classes. However, a common limitation across these methods is the neglect of inter-class relationships particularly how semantic similarity between classes influences misclassification patterns.

Our empirical analysis reveals that misclassifications of minority classes often occur in favor of semantically similar majority classes. To capture and leverage these subtle relationships, we introduce the Relation Contextual Module (RCM). Unlike general class imbalance metrics that can be inferred from data distributions, inter-class dynamics must be learned from the model's behavior during training. RCM facilitates this by generating a confusion matrix at each training batch, comparing the model's predictions with ground-truth labels. This matrix is normalized with respect to the ground truth, enabling a dynamic estimation of misclassification biases between specific class pairs. To ensure stability

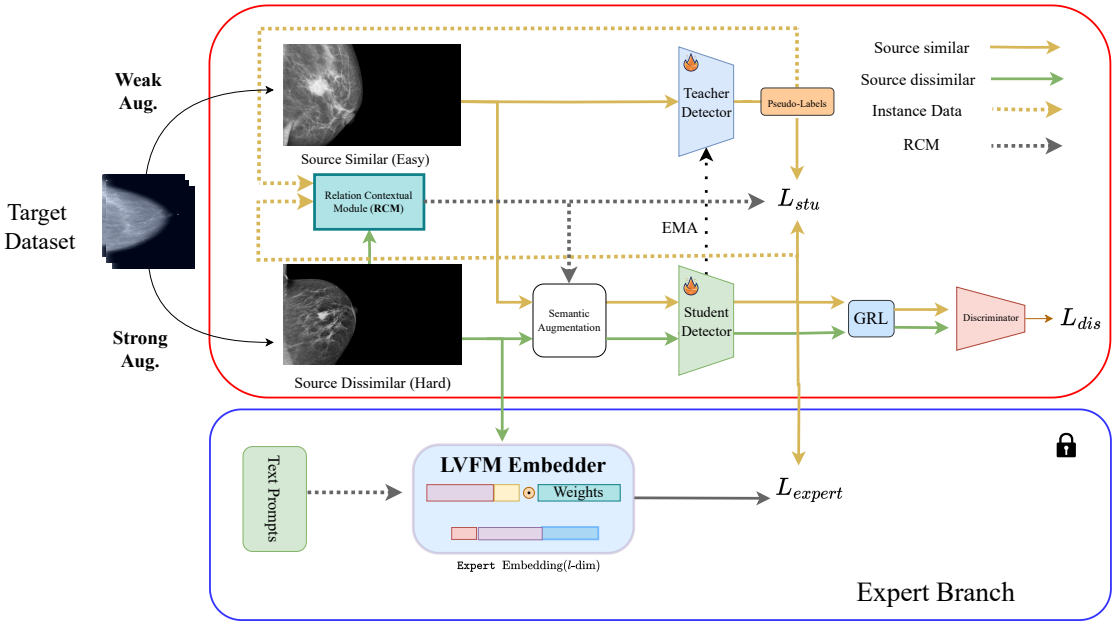

Fig. 2: ***Detail architecture of (GT):*** *The Relation-Contextual Module (RCM) computes a confusion matrix updated via EMA, feeding both the augmentation policy (SA) and the reweighting module (SAL). Training employs pseudo-labels from the teacher and optional Expert guidance. During inference, the Expert and RCM branches are removed, yielding identical computational cost to the baseline detector.*

and robustness over time, we apply an Exponential Moving Average (`EMA`) to iteratively update a global confusion matrix. This evolving matrix serves as a reliable representation of the model's long-term class bias patterns. Beyond providing smoother updates, the use of `EMA` eliminates the requirement for all classes to be present in every batch—thus streamlining the learning process. The procedure for updating this matrix is detailed in Algorithm 1.

When applied to target domain samples, `RCM` relies on high-confidence pseudo-labels to estimate class biases in the absence of ground-truth annotations, offering a reliable approximation that mirrors the supervised setup on the source domain.

Aligning source and target domains is common in traditional domain adaptation tasks [58, 152, 116]. While alignment can be performed in either data space [12] or feature space [99], the absence of source data presents a unique challenge for domain alignment.

Even though the source data is inaccessible, the source-pretrained model retains crucial knowledge about the source domain. Following [21], we partition the target dataset into two groups, leveraging the pre-trained model to establish an explicit source-target domain distinction. To achieve this, a detection variance-based criterion is used, where variances are computed from the pre-trained model's predictions on the target samples. A higher variance suggests a stronger resemblance to the source domain [21]. In particular, the model exhibits greater uncertainty (hard samples) when predicting source-similar images, leading to elevated variance values compared to source-dissimilar images. The detection variance is determined using the following formulation:

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

where $F_{\theta_s}(X_i)$ are the predictions of image X_i via the source-pretrained model. Since this calculation is computationally intractable, we approximate it with Monte-Carlo sampling using dropout, following the method proposed by Gal and Ghahramani [35]. This approximation is achieved by conducting M stochastic forward passes while keeping the detection model unchanged [9].

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:

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

$$v_{ci} = \frac{1}{MN_i} \sum_{j=1}^{N_i} \sum_{m=1}^{M} \|\mathbf{c_{ij}^m} - \bar{\mathbf{c}}_{\mathbf{ij}}\|^2, \tag{4}$$

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 $vl_i \geqslant \sigma$, we categorize X_i as source-similar; otherwise, it is deemed source-dissimilar, with $\sigma \in (0, 1)$ being a predefined threshold. This process effectively partitions the target domain into source-similar and source-dissimilar subsets for query-based adversarial alignment between domains.

Oversampling is a widely used strategy in classification tasks to mitigate class imbalance by increasing the representation of minority classes. However, this method presents challenges in object detection, where individual images often contain multiple object classes. Our analysis of breast cancer detection datasets such as DDSM [66] reveals that most images do not consistently feature even a single instance of majority classes, making conventional image-level resampling approaches largely ineffective. This calls for more targeted, instance-aware augmentation techniques tailored to object detection.

Following the instance-level augmentation strategy proposed by Zhang et al. [140], we extract object instances using bounding box annotations and strategically insert them into other images. To guide this process, we leverage our Relation Contextual Module (`RCM`) to identify majority and minority classes based on their likelihood of correct classification.

We compute the average diagonal value of the `RCM` —representing the mean self-classification probability—using the formula:

$$\texttt{RCM}_{\text{avg}} = \frac{1}{C} \sum_{c=0}^{C} \texttt{RCM}(c, c), \tag{5}$$

where C is the number of classes. Classes with $\texttt{RCM}(c, c)$ values above this mean are labeled as majority classes, while those below are considered minority.

Unlike prior methods that randomly overlay instances, we employ a relation-guided MixUp strategy. For each base instance, we select a paired instance through weighted random sampling, where the weights are determined by inter-class misclassification probabilities captured in the `RCM`. For majority base instances, we use the column vector $\texttt{RCM}(:, c)$, setting $\texttt{RCM}(c, c) = 0$ to exclude self-augmentation and focus on classes often confused with the majority. For minority base instances, the row vector $\texttt{RCM}(c, :)$ is used without masking, allowing for beneficial self-augmentation.

Paired instances are resized to match the base instance for bounding box consistency. MixUp is then applied as:

$$\begin{aligned} \hat{I} &= \beta \cdot I_{\text{base}} + (1 - \beta) \cdot I_{\text{mix}}, \\ \hat{c} &= \beta \cdot c_{\text{base}} + (1 - \beta) \cdot c_{\text{mix}}, \end{aligned} \tag{6}$$

where I_{base} and I_{mix} are the cropped images of the base and mixed instances, and $\hat{I}$ is the resulting augmented image. Correspondingly, c_{base} and c_{mix} denote the respective class vectors, with $\hat{c}$ being the interpolated class label.

To support domain adaptation, we apply this augmentation to both source similar and dissimilar domains. For source-like images, we sample instances from both domains to promote knowledge transfer. For source dissimilar images, we prioritize dissimilar domain instances, falling back to source similar instances only when necessary. Importantly, we refrain from augmenting minority base instances in the target domain to preserve their semantic integrity, helping the model maintain focus on real target distributions and avoid domain drift.

To enable inter-semantic Augmentation, we maintain a dynamic repository of class-specific

Algorithm 1 Relation Contextual Module (RCM) Update Procedure

Require: Global class-relation matrix $\mathbf{R} \in \mathbb{R}^{C\times C}$ initialized to zero; `EMA` decay factor $\beta \in [0, 1]$.

while training **do**
  Initialize batch-specific relation matrix $\mathbf{R}_b \in \mathbb{R}^{C\times C}$ to zero.
  for each training example with ground-truth label c_i and predicted label x_i **do**
    $\mathbf{R}_b[c_i, x_i] \leftarrow \mathbf{R}_b[c_i, x_i] + 1$
  end for
  for each class $c \in \{1, \ldots, C\}$ **do**
    Normalize $\mathbf{R}_b[c, :]$ to sum to 1.
    Update $\mathbf{R}[c, :] \leftarrow \beta \cdot \mathbf{R}[c, :] + (1-\beta) \cdot \mathbf{R}_b[c, :]$
  end for
end while

instance crops, referred to as the *Cropbank* [140]. These instance crops are extracted from bounding box annotations in labeled source images as well as pseudo-labeled target images. To facilitate domain-specific augmentation, separate Cropbanks are maintained for source similar and source dissimilar subsets. Each Cropbank is updated in a first-in-first-out (`FIFO`) manner to ensure sample diversity and freshness. This strategy is particularly advantageous for the source dissimilar Cropbank, where early-stage pseudo-labels may be less reliable; replacing them over time helps improve augmentation quality as training progresses.

We use the source-/target-similarity split only to route instance crops to the corresponding Cropbank; the `RCM` iteslf remains a single global estimator shared across subsets. The same confusion statistics condition the relation-guided augmentation (SA) and the Semantic-Aware Loss (SAL) across both subsets.

4.2 Semantic-Aware Loss

To further mitigate class bias, we introduce a weighted parameter into the classification loss, guided by the Relation Contextual Module (`RCM`) for foreground classes. This weighting scheme emphasizes classes that are frequently misclassified as dominant majority classes, thereby directing the model's attention toward improving performance on underrepresented or confused categories.

To accentuate this focus, we apply a non-linear transformation to the `RCM` values as follows:

$$w_i = \begin{cases} \sqrt{1 - \texttt{RCM}(c_i, x_i)}, & \text{if } c_i = x_i \\ \sqrt{\frac{\texttt{RCM}(c_i, x_i)}{\texttt{RCM}(c_i, c_i)}}, & \text{otherwise} \end{cases} \tag{7}$$

where w_i denotes the weight for the i^{th} instance, c_i is its pseudo-label, and x_i is the predicted class. When $c_i \neq x_i$, we normalize using the diagonal entry $\texttt{RCM}(c_i, c_i)$ to scale the misclassification likelihood relative to the class's overall performance.

Background class weights are uniformly set to 1 to prevent model bias against them. To reconcile the disparity between foreground and background weights, we normalize foreground instance weights so that their mean equals that of the background class:

$$W_f = \frac{W_f}{\text{mean}(W_f)}, \tag{8}$$

where W_f represents the set of weights for foreground instances.

To avoid excessive influence from extreme weights, we apply a regularization term λ_l to all class-relation weights:

$$W = \frac{W + \lambda_l}{1 + \lambda_l}, \tag{9}$$

ensuring smoother gradients and preventing the loss from being overly sensitive to outlier weights.

The weighted classification loss is then defined as:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot \texttt{CE}(c_i, x_i), \tag{10}$$

where N is the total number of instances, and $\texttt{CE}(\cdot)$ denotes the standard cross-entropy loss.

Because we regularize instance weights as per Eq. (9), the post-regularization weights satisfy

$$W \in \left[\frac{\lambda_\ell}{1 + \lambda_\ell}, 1 \right] \quad \text{whenever } W \in [0, 1].$$

Consequently, the SAL $\mathcal{L}_{\text{cls}}$ is pointwise bounded, preventing gradient spikes from rare but extreme

confusions. In ablations, removing this regularizer reduces performance casuing a notable drop(e.g., 50.8 → 47.8 mAP on Cityscapes→Foggy), indicating that boundedness is essential to avoid EMA-amplified teacher drift (see §4.2 and Appendix. A).

4.3 Grounded Supervision

We introduce a framework that leverages Large Vision Foundation Models (`LVFMs`) through expert distillation, as depicted in 2. We introduce an Expert branch aimed at assimilating the strengths of `LVFMs` into a unified learning approach. As an initial assumption, we posit that the expert model can effectively represent a wide variety of images sourced from datasets like `ImageNet1K` or `ImageNet21k` [26], `LAION-400M` [103], or `DataComp1B` [34]. We have considered many prominent families of teacher models: `CLIP` [90], `DINOV2` [87], `ViT` [32]and `SAM` [64], `GroundingDINO` [78], and `BioMedParse` [146]. These models have exhibited outstanding performance across a wide array of tasks, with `CLIP` demonstrating broad proficiency, `DINOV2` excelling in downstream dense tasks such as semantic segmentation under linear probe, and `SAM` particularly shining in open-vocabulary segmentation tasks. The newly introduced Expert branch incorporates the pretrained `LVFMs`. In the context of knowledge distillation, utilizing a single expert's embedding directly for distilling knowledge to the student model may not suffice. We found that relying on an expert's embedding directly for distilling knowledge to the student model is insufficient for effective pseudo-label correction during adaptation. Hence, we propose using expert predictions and introducing a pseudo-supervised loss to guide the student model out from the local minima.

Our objective is to transfer the expertise of the expert model to the student model. The expert model $E(\cdot)$ is frozen, while the student model $S(\cdot)$ is updated via gradient descent during adaptation. To supervise the student using pseudo labels from the expert, we employ a combination of a consistency loss (mean teacher loss) and a bounding box regression loss.

Let x_i be an input image. The expert produces pseudo-labels in the form of bounding box predictions $E^i_{\text{bbox}} = E(x_i)$, which are used to supervise the student's prediction $S^i_{\text{bbox}} = S(x_i)$. We define the loss as:

$$\begin{aligned}\mathcal{L}_{\text{expert}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}(S^i_{\text{bbox}}, E^i_{\text{bbox}}) \\ + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}}(S^i_{\text{bbox}}, E^i_{\text{bbox}})\end{aligned} \tag{11}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss (semantic aware loss), and $\mathcal{L}_{\text{reg}}$ is the bounding box regression loss (e.g., smooth `L1` or `GIoU` loss). The weights λ_{cls} and λ_{reg} control the contribution of each term.

Zero-shot vs. training-time guidance. We benchmark the expert alone (frozen LVFM zero-shot detector), student–teacher without Expert (pseudo-label self-training), and student–teacher with Expert (our default). Across both natural and medical domains, the Expert-only zero-shot detector underperforms supervised/self-trained detectors, while training-time Expert guidance consistently improves over pseudo-label self-training by correcting class confusions and stabilizing teacher updates (Section §4.2, Table A1).

For natural scenes, GroundingDINO + SAM yields the strongest guidance due to robust spatial grounding, while for medical transfers, CLIP-ViT with domain-tuned text prompts performs best. In both cases we observe CLIP + DINOv2 ensembling offers small but consistent gains. All variants use the same expert-supervision loss (Eq. (11)) without extra hyperparameters or architectural changes ((Table A2).

This unified loss ensures that the student model learns both semantic consistency and localization accuracy from the expert model via pseudo-supervision.

4.4 Overall Objective and Training Strategy.

The final objective combines supervised, unsupervised, and discriminator losses:

$$\mathcal{L} = \lambda_u \mathcal{L}_{\text{stu}} + \lambda_d \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{Expert}}, \tag{12}$$

where λ_u and λ_d control the contributions of the unsupervised and discriminator losses, respectively, and $\mathcal{L}_{\text{stu}}$, and $\mathcal{L}_{\text{Expert}}$ have been formulated in Eq. (1), and (11) respectively. As the teaching process continues, an adequate amount of pseudo-boxes will be produced, and the influence of λ_u should be lowered to avoid the encoder over-fitting to pseudo-labels. As a result, We decay λ_d as adaptation continues.

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
DDSM to INBreast	UMT [137]	CVPR'21	✗	0.01	0.09	0.15	0.21	0.204	0.319
	SFA [114]	MM'21	✗	0.03	0.13	0.23	0.35	0.241	0.338
	H2FA [126]	CVPR'22	✗	0.13	0.32	0.45	0.52	0.575	0.312
	AQT [53]	IJCAI'22	✗	0.11	0.37	0.44	0.57	0.680	0.349
	AT [71]	CVPR'22	✗	0.19	0.38	0.65	0.75	0.721	0.512
	D-Adapt [55]	ICLR'22	✗	0.11	0.29	0.45	0.59	0.731	0.414
	HT [29]	CVPR'23	✗	0.17	0.49	0.61	0.69	0.704	0.362
	CLIPGAP [109]	CVPR'23	✗	0.15	0.55	0.61	0.75	0.712	0.445
	ConfMIX [84]	WACV'23	✗	0.19	0.47	0.58	0.73	0.737	0.409
	MRT [148]	ICCV'23	✗	0.16	0.54	0.64	0.72	0.789	0.489
	D-MASTER [6]	MICCAI'24	✗	0.25	0.61	0.70	0.82	0.808	0.524
	Mexforer [114]	MM'21	✓	0.02	0.03	0.03	0.03	0.060	0.090
	IRG [27]	CVPR'23	✓	0.05	0.05	0.07	0.09	0.110	0.120
	GT (Ours)		✓	**0.06**	**0.45**	**0.65**	**0.92**	**0.589**	**0.758**

Table 1: Results of adaptation from DDSM to INBreast.

The Expert branch operates only during training to provide supervision and is removed during inference. So, the deployed detector is identical to the student model, ensuring that inference complexity and computational cost remain the same as the baseline configuration. Additional implementation and runtime details are provided in §4.3.

5 Experiments

We conduct extensive experiments to assess the effectiveness of our approach across various challenging medical imaging datasets, as well as standard UDA and SFOD benchmarks. In the UDA setting, both source and target domain data are accessible during training, facilitating adaptation. Conversely, in the SFOD scenario, adaptation is performed using only a source-trained model without access to source domain data. Additionally, we perform ablation studies employing diverse exchange strategies to validate the efficacy of our proposed methods. We further analyze the promising results obtained by our method through comprehensive visualizations and component-wise evaluations.

5.1 Comparison with Current SOTA Methods

We evaluate the performance of our proposed Grounded Teacher (GT) method against other approaches on all three medical benchmarks and the natural benchmark mentioned earlier for generalizability. Since UDA and SFDOAD share similar task settings, we conducted comparisons with both. Table 1, Table 2 and Table 3 present the comparison results on medical image datasets. Table 5 presents comparison of the natural dataset. Our proposed GT consistently outperforms existing all SOTA methods, demonstrating generalizability and significant improvements across both natural and medical settings.

DDSM to INBreast. Adaptation from large to small-scale medical datasets with different modalities. Here, we consider DDSM [66] dataset as the source domain and INBreast [85] as the target domain. Results are presented in Table 1. Our proposed method demonstrates superior performance across various False positives per Image (FPI) values compared to existing methods as displayed.

RSNA to INBreast. Adaptation across medical datasets with different machine-acquisitions. This is vital for enhancing healthcare outcomes, improving diagnostic accuracy, and facilitating better clinical decisions. To evaluate our method's performance, we adapt a model trained on DDSM [66] to RSNA [11]. Results for all FPI values are presented in Table 2, demonstrating that our method achieves state-of-the-art performance on this benchmark.

DDSM to RSNA. This experiment evaluates domain adaptation across datasets collected from distinct geographical regions. Specifically, we consider the DDSM [66] dataset as the source domain and RSNA [11] as the target domain. As shown in

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
DDSM to RSNA	D-Adapt [55]	ICLR'21	✗	0.04	0.12	0.18	0.29	0.439	0.263
	AT [73]	CVPR'22	✗	0.16	0.28	0.35	0.42	0.486	0.338
	H2FA [127]	CVPR'22	✗	0.03	0.13	0.18	0.36	0.634	0.236
	MRT [147]	ICCV'23	✗	0.32	0.52	0.69	0.72	0.741	0.352
	Mexforer [114]	MM'21	✓	**0.24**	0.31	0.39	0.39	0.336	0.287
	IRG [27]	CVPR'23	✓	0.16	0.25	0.37	0.39	0.308	0.235
	GT (Ours)		✓	0.10	**0.43**	**0.58**	**0.91**	**0.781**	**0.530**

Table 2: Results of adaptation from DDSM to RSNA.

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
RSNA to INBreast	D-Adapt [55]	ICLR'21	✗	0.00	0.06	0.09	0.10	0.381	0.362
	AT [73]	CVPR'22	✗	0.01	0.08	0.10	0.15	0.385	0.311
	H2FA [127]	CVPR'22	✗	0.02	0.08	0.10	0.12	0.483	0.315
	MRT [147]	ICCV'23	✗	0.03	0.09	0.12	0.17	0.739	0.587
	Mexforer [114]	MM'21	✓	0.02	0.03	0.03	0.03	0.060	0.090
	IRG [27]	CVPR'23	✓	**0.05**	0.05	0.07	0.09	0.110	0.120
	GT (Ours)		✓	0.01	**0.28**	**0.49**	**0.90**	**0.638**	**0.605**

Table 3: Results of adaptation from RSNA to INBreast.

Table 2, our proposed method consistently outperforms existing approaches across all FPI thresholds, demonstrating its robustness and effectiveness in cross-domain generalization.

DDSM to DeepLesion (CT) We follow a source-free protocol on DeepLesion by splitting scanner/protocols to induce shift and training uses the same hyperparameters as other medical runs. We report AP50 and sensitivity at fixed FPPI. GT outperforms strong SFOD baselines by a meaningful margin while maintaining box-level consistency. Full protocol and curves are in Appx M (This complements mammography and CXR, demonstrating generalization across imaging physics and anatomy).

Cityscapes to Foggy Cityscapes. Object detectors often experience a significant drop in performance when deployed under adverse real-world conditions such as fog, due to the domain shift caused by the lack of such conditions in the training data. Domain adaptation aims to bridge this gap between normal and adverse weather scenarios. To investigate this, we conduct experiments on the widely-used Cityscapes → FoggyCityscapes benchmark. As presented in Table 5, student-teacher based frameworks consistently outperform non-student-teacher approaches by a notable margin.

Under identical VGG-16 backbones and standard Cityscapes→Foggy settings, our method GT reaches 50.8 mAP, outperforming contemporary SF-UT and LPLD baselines. Furthermore, it demonstrates notable gains in minority object classes as rider improves by +6.3 AP50 and bicycle by +8.9 AP50 over the best prior baselines, tracking SAL's emphasis on minority→majority confusion and SA's context diversification. This highlights its robustness and adaptability in challenging weather conditions, and supports that confusion-guided SA/SAL with training-only Expert guidance address minority-class instability beyond what strong/weak augmentation and low-confidence distillation alone provide.

5.2 Compute Overhead

All experiments here were conducted on a single NVIDIA A100 (40 GB) GPU using identical data-loading and optimization settings as the baseline mean-teacher detector to ensure fair measurement. Integrating RCM + SAL introduces only 2–3 % additional training time and negligible

Exp	Method	Venue	SF	R@0.05	R@0.3	R@0.5	R@1.0	AUC	F1-score
DDSM to DeepLesion	D-Adapt [55]	ICLR'21	✗	0.04	0.12	0.18	0.29	0.439	0.263
	AT [73]	CVPR'22	✗	0.16	0.28	0.35	0.42	0.486	0.338
	H2FA [127]	CVPR'22	✗	0.03	0.13	0.18	0.36	0.634	0.236
	MRT [147]	ICCV'23	✗	0.32	0.52	0.69	0.72	0.741	0.352
	Mexforer [114]	MM'21	✓	**0.24**	0.31	0.39	0.39	0.336	0.287
	IRG [27]	CVPR'23	✓	0.16	0.25	0.37	0.39	0.308	0.235
	GT (Ours)		✓	0.10	**0.43**	**0.58**	**0.91**	**0.781**	**0.530**

Table 4: Results of adaptation from DDSM to DeepLesion.

Method	Venue	SF	Person	Rider	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
DA-Faster [18]	CVPR'18	✗	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
EPM [49]	ECCV'20	✗	44.2	46.6	58.5	24.8	45.2	29.1	28.6	34.6	39.0
SSAL [86]	NIPS'21	✗	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
SFA [114]	MM'21	✗	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
UMT [28]	CVPR'21	✗	33.0	46.7	48.6	34.1	56.5	46.8	30.4	37.3	41.7
D-adapt [55]	ICLR'21	✗	40.8	47.1	57.5	33.5	46.9	41.4	33.6	43.0	43.0
TIA [144]	CVPR'22	✗	34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
PT [44]	ICML'22	✗	40.2	48.8	63.4	30.7	51.8	30.6	35.4	44.5	42.7
MTTrans [136]	ECCV'22	✗	47.7	49.9	65.2	25.8	45.9	33.8	32.6	46.5	43.4
SIGMA [68]	CVPR'22	✗	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
O2net [41]	MM'22	✗	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
AQT [53]	IJCAI'22	✗	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
AT [71]	CVPR'22	✗	43.7	54.1	62.3	31.9	54.4	49.3	35.2	47.9	47.4
TDD [45]	CVPR'22	✗	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	49.2
CIGAR [79]	CVPR'23	✗	45.3	45.3	61.6	32.1	50.0	51.0	31.9	40.4	44.7
CSDA [38]	ICCV'23	✗	46.6	46.3	63.1	28.1	56.3	53.7	33.1	39.1	45.8
HT [30]	CVPR'23	✗	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
MRT [147]	ICCV'23	✗	52.8	51.7	68.7	35.9	58.1	54.5	41.0	47.1	51.2
SFOD [69]	AAAI'21	✓	21.7	44.0	40.4	32.6	11.8	25.3	34.5	34.3	30.6
SFOD-Mosaic [69]	AAAI'21	✓	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5
HCL [52]	NIPS'21	✓	38.7	46.0	47.9	33.0	45.7	38.9	32.8	34.9	39.7
SOAP [124]	IJIS'21	✓	35.9	45.0	48.4	23.9	37.2	24.3	31.8	37.9	35.5
LODS [67]	CVPR'22	✓	34.0	45.7	48.8	27.3	39.7	19.6	33.2	37.8	35.8
AASFOD [21]	AAAI'23	✓	32.3	44.1	44.6	28.1	34.3	29.0	31.8	38.9	35.4
IRG [110]	CVPR'23	✓	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
PETS [77]	ICCV'23	✓	**46.1**	52.8	**63.4**	21.8	46.7	5.5	37.4	48.4	40.3
DACA [145]	IJCV'24	✓	44.7	31.2	60.1	**53.1**	53.9	0.0	27.8	45.9	39.9
LPLD [133]	ECCV'24	✓	39.7	49.1	56.6	29.6	46.3	26.4	36.1	43.6	40.9
SF-UT [43]	ECCV'24	✓	40.9	48.0	58.9	29.6	51.9	50.2	36.2	44.1	45.0
GT (Ours)		✓	42.7	**55.4**	61.7	40.7	**62.0**	**54.6**	**39.1**	**53.0**	**50.8**
Oracle		✗	66.3	61.1	80.8	45.6	68.8	52.0	49.1	54.9	59.8

Table 5: Results of adaptation from Normal Cityscapes to Foggy weather Cityscapes (CF). "SF" refers to source-free setting. "Oracle" refers to the models trained by using labels during training.

VRAM increase (¡ 0.3 GB). Enabling the training-only Expert branch for extracting frozen LVFM features adds 10–12 % training time and about 1 GB of VRAM usage. Importantly, these modules are disabled during inference and the Inference cost remains unchanged as the deployed student detector has the same architecture, latency, and FLOPs as the baselines. Additional details about profiling logs, GPU-memory traces, and wall-clock analyses are provided in Appendix C.

Method	RCM	SA	SAL	mAP
Base (AT [72])				30.4
+ SA	✓	✓		41.8
+ SAL	✓		✓	44.2
GT (Full)	✓	✓	✓	50.8

Table 6: **Ablation studies on GT components.** RCM is included in all studies as it forms the basis of CRA and SAL. We report the mean average precision at 0.50 IoU (mAP). Our contributions are not included in the base framework (AT [71]).

Selection Method	mAP
Random (0.5)	46.4
Semantic Augmentation (1.0)	47.5
Semantic Augmentation (0.5)	50.8

Table 7: Comparison of selecting class instances randomly and via CRA. Values in brackets refer to the likelihood of an instance being augmented.

5.3 Ablation Studies

To verify the significance of our contributions, we conducted an ablation study. All experiments within this study were performed on the Cityscapes → Foggy Cityscapes benchmark using the VGG16 backbone.

5.3.1 Quantitative Ablation

Table 6 presents a quantitative analysis of the impact of each component within our framework. The base framework prior to integrating our modules aligns with the AT model [71], achieving a 30.4 mAP. Given that the Relation-Contextual Module (RCM) is integral to both class-relation augmentation and semantic-aware loss, it remains a constant across all experimental variations. Integrating Semantic Augmentation (SA) along with RCM increases the performance to 41.8 mAP, a gain of +11.4 over the base. Adding the Semantic Aware Loss (SAL) with RCM separately results in a higher mAP of 44.2, demonstrating a +13.8 improvement compared to the baseline, indicating that SAL provides a greater performance boost than SA in this ablation setting. Notably, Class-Relation Augmentation (CRA) significantly reduces the performance disparity between minority and majority classes, as evidenced. Finally, our full GT model, which combines RCM, SA, and SAL, achieves the highest score of 50.8 mAP, which not only boosts overall performance but also underscores our method's efficacy in managing class imbalance.

5.3.2 Impact of Augmentation

In our study, we assess the influence of augmentation strategies, focusing on both the proportion of augmented data and the criteria for selecting instances to augment. Image augmentation plays a pivotal role in our methodology, enhancing the dataset by incorporating additional representations, particularly of minority classes. However, excessive augmentation can lead to the model learning less accurate class representations. To mitigate this, we judiciously augment a random subset of images.

The pairing of class instances during augmentation is crucial for improving its effectiveness. Unlike random pairing, our Semantic Augmentation (SA) method prioritizes pairing instances from closely related minority and majority classes. This targeted approach ensures that the Mixup outputs are more meaningful, thereby enhancing the performance of minority classes.

Table 7 presents the experimental results of this approach. We compare the effects of randomly selecting class instances for Mixup against our SA method across different augmentation probabilities, which represent the likelihood of an instance in the base image being augmented. Applying Mixup randomly yields a modest improvement of +1.1 mAP in overall performance. In contrast, employing SA leads to a more substantial increase of +4.4 mAP. These findings indicate that pairing highly related classes during Mixup enhances minority class performance with minimal impact on majority classes. Furthermore, our experiments reveal that excessive augmentation can negatively affect model performance, underscoring the importance of a balanced augmentation strategy.

5.3.3 Weighing Strategy for Semantic Loss

To enhance the performance of minority classes, we introduce a Semantic Aware Loss (SAL) function, as defined in Eq 10. Class-level loss functions

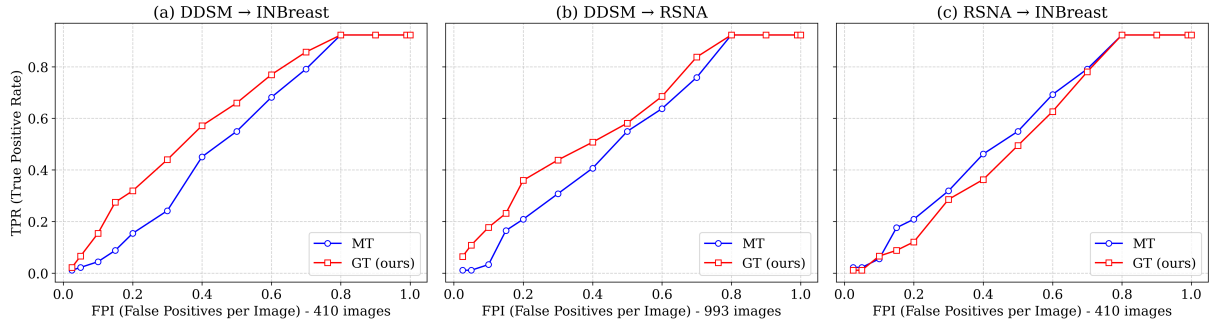

Fig. 3: **FROC analysis of model generalization on unseen target domains**. Performance comparison between our method (`GT`) and the MT baseline for detection tasks under domain shift: (a) DDSM→INBreast (410 images), (b) DDSM→RSNA (993 images), and (c) RSNA→INBreast (410 images), where our method demonstrates superior performance across different domain shifts. Higher curves indicate better trade-offs between sensitivity (`TPR`) and false positives per image (`FPI`).

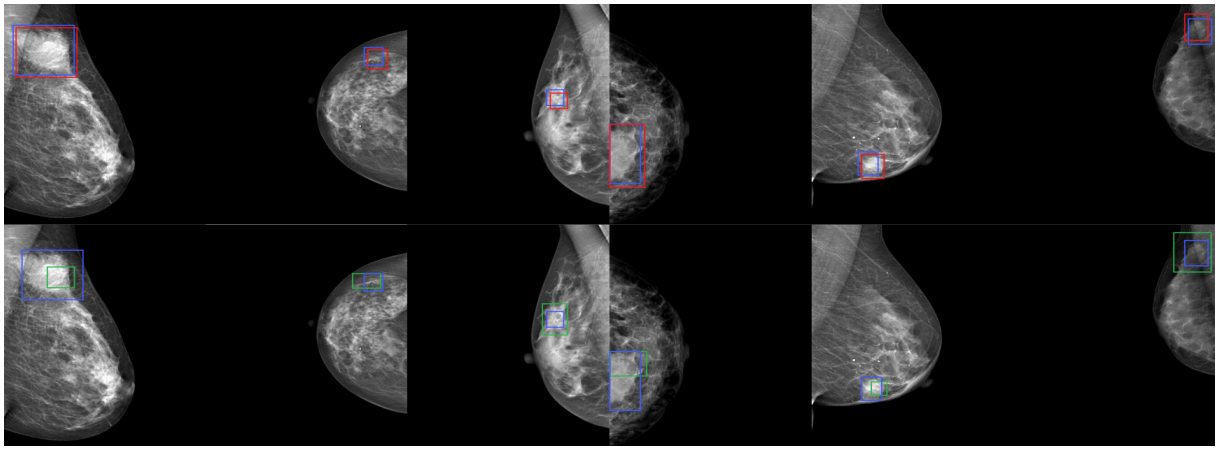

Fig. 4: Qualitative results of our proposed Grounded Teacher on Medical settings (DDSM → INBreast). We present visual comparisons between `AT` (top row) and `GT` (bottom row). Our method, `GT`, effectively identifies Malignancies better and mitigates false negatives. Bounding box colors indicate: Blue — Ground truth, Green — true positive, and Red — false negative.

are a prevalent strategy for addressing dataset class imbalances. We compare our `SAL` with a variant of existing class-level losses that utilize only the diagonal elements of the Inter-Class Relation module, which correspond to the ground-truth class likelihood for accurate classification. In contrast, `SAL` leverages the likelihood of the ground-truth being classified as the predicted class to influence its loss function.

As shown in Table 8, incorporating inter-class relationships through `SAL` achieves an mAP of 50.8, a performance improvement of +3.7 `mAP` compared to the base. While `SAL` effectively utilizes inter-class information, there are instances where it might excessively penalize well-performing classes without proper constraints. To mitigate this, we apply a regularization term, as specified in Eq 9. Omitting this regularization (`SAL` w/o Reg.) results in an `mAP` of 47.8, demonstrating a significant performance drop of -3.0 `mAP` compared to the full `SAL`. This clearly highlights the importance of the regularization term, likely preventing certain class weights from becoming disproportionately small during training.

5.3.4 Qualitative Results

Figure 5 presents the qualitative outcomes of our method, comparing predictions from baseline

Fig. 5: Qualitative results of our proposed Grounded Teacher on Natural settings (Cityscapes → Foggy Cityscapes). We present visual comparisons between `AT` (top row) and `GT` (bottom row). Our method, `GT`, effectively mitigates misclassifications, reduces false negatives, and eliminates false positives. Bounding box colors indicate: Green — true positive, Blue — misclassified, Red — false negative, and Pink — false positive.

Class Weight Strategy	mAP
Class-Level	47.1
`SAL` w/o Reg.	47.8
`SAL`	50.8

Table 8: Class loss weighting strategies. *Class-Level* uses only the diagonal values in our `RCM` with regularization. ***SAL*** refers to our full Semantic Aware Loss.

`AT` [71] in the top row with those from the Ground Truth (`GT`) in the bottom row. The `GT` effectively corrects misclassifications, as highlighted by the blue boxes. For instance, in columns 1, 2, and 4, the `GT` accurately labels 'person', 'car', and 'truck', respectively, rectifying the `AT's` errors. Moreover, the `GT` demonstrates a notable reduction in both false positives and false negatives, indicated by the pink and red boxes, respectively. This enhancement signifies improved detection accuracy across various object scales and classes. This improvement is particularly evident in column 3, where the `GT` successfully identifies small-scale objects and provides more accurate detections for larger-scale objects. Also in Fig 5, we show how qualitatively the predictions are improved with various proposed modules.

References

[1] Zeynettin Akkus, Alsu Galimzianova, Assaf Hoogi, Daniel L. Rubin, and Bradley J. Erickson. Deep learning for brain mri segmentation: State of the art and future directions. *Journal of Digital Imaging*, 30(4):449–459, 2017.

[2] Yazeed Alashban. Breast cancer detection and classification with digital breast tomosynthesis: a two-stage deep learning approach. *Diagnostic and Interventional Radiology*, 2024.

[3] Munsif Ali, Leonardo Rossi, and Massimo Bertozzi. Cfts-gan: Continual few-shot teacher student for generative adversarial networks, 2024.

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

[5] Tajamul Ashraf and Tisha Madame. Federated learning framework for x-ray imaging. In *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care: First Deep Breast Workshop, Deep-Breath 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 10, 2024, Proceedings*, page 13. Springer Nature, 2024.

[6] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199. Springer, 2024.

[7] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings*

Table 9: **Ablation study isolating the contribution of Expert guidance.** We compare a frozen LVFM detector alone (Zero-shot Expert), self-training without Expert supervision (No-Expert), and the full GT with both pseudo-labels and Expert guidance (PL+Expert). Results are AP_{50} for medical transfers and mAP for Cityscapes→Foggy.

Method Variant	DDSM→INBreast	DDSM→RSNA	RSNA→INBreast	Cityscapes→Foggy
Zero-shot Expert (frozen LVFM only)	68.4	70.2	65.7	42.1
No-Expert (PL only)	74.6	75.3	71.9	47.8
GT (PL + Expert)	**78.9**	**79.5**	**76.1**	**50.8**

Table 10: **Ablation of different Large-Vision-Foundation-Model (LVFM) experts.** Each expert supervises the student during training using the same loss in Eq. (11). Best results for each task are highlighted in bold.

Expert Type	DDSM→INBreast	DDSM→RSNA	RSNA→INBreast	Cityscapes→Foggy
CLIP-ViT (prompt-tuned)	**78.9**	**79.5**	**76.1**	49.2
DINOv2 (self-sup.)	77.3	77.9	74.6	48.6
GroundingDINO + SAM	76.2	76.7	75.3	**50.8**
CLIP + DINOv2 (ensemble)	78.1	78.8	75.9	50.2

Table 11: **Evaluation on the DeepLesion dataset under source-free setting.** Source and target splits are formed by scanner/protocol differences. Metrics: AP_{50} and sensitivity at 4 false positives per image (FPPI).

Method	AP_{50}	Sensitivity@4FPPI
SF-UT (ECCV'24)	64.8	75.2
LPLD (ECCV'24)	65.7	76.5
GT (ours)	**69.9**	**80.3**

of the IEEE conference on computer vision and pattern recognition, pages 1081–1089, 2015.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12576–12585, 2021.

[14] Qingchao Chen and Yang Liu. Structure-aware feature fusion for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10567–10574, 2020.

[15] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7976–7985, 2018.

[16] 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.

[17] 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.

[18] 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)*, June 2018.

[19] 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.

[20] H. Chou, S. Chang, J. Pan, W. Wei, and D. Juan. Remix: rebalanced mixup. *Computer Vision – ECCV 2020 Workshops*, pages 95–110, 2020.

[21] 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*, volume 37, pages 452–460, 2023.

[22] 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.

[23] 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.

[24] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[25] 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.

[26] 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.

[27] 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.

[28] 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, June 2021.

[29] 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 (CVPR)*, pages 23829–23838, June 2023.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, pages 1050–1059, 2016.

[36] A. Galdrán, G. Carneiro, and M. Á. G. Ballester. Balanced-mixup for highly imbalanced medical image classification. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 323–333, 2021.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] Yan Hao, Florent Forest, and Olga Fink. Simplifying source-free domain adaptation for object detection: Effective self-training strategies and performance insights. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024.

[44] 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.

[45] 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, June 2022.

[46] 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.

[47] 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 *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[48] 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.

[49] 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.

[50] Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised Domain Adaptation with Imbalanced Cross-Domain Data. In *IEEE/CVF International Conference on Computer Vision*, pages 4121–4129, December 2015.

[51] 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.

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

[53] 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.

[54] 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.

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

[56] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation. In *International Conference on Machine Learning*, pages 4816–4827, November 2020.

[57] 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.

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

[59] S.R. Kebede, F.G. Waldamichael, T.G. Debelee, and et al. Dual view deep learning for enhanced breast cancer screening using mammography. *Sci Rep*, 14:3839, 2024.

[60] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T. Tan. CAT: Exploiting Inter-Class Dynamics for Domain Adaptive Object Detection, 2024.

[61] A. Khalid, A. Mehmood, A. Alabrah, B. F. Alkhamees, F. Amin, H. AlSalman, and G. S. Choi. Breast cancer detection and prevention using machine learning. *Diagnostics (Basel)*, 13(19):3113, Oct 2023.

[62] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[63] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[64] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[65] 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.

[66] 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.

[67] 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.

[68] 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, June 2022.

[69] 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.

[70] 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*, volume 35, pages 8474–8481, 2021.

[71] 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.

[72] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7571–7580, 2022.

[73] 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.

[74] 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.

[75] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[76] 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.

[77] 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.

[78] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang. Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 423–433, 2023.

[85] 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.

[86] 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.

[87] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[88] 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.

[89] Oscar Pina and Verónica Vilaplana. Unsupervised domain adaptation for multi-stain cell detection in breast cancer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5066–5074, 2024.

[90] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, page 91–99, 2015.

[96] 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.

[97] F. Ryan, K. L. Román, B. Z. Gerbolés, K. M. Rebescher, M. S. Txurio, R. C. Ugarte, M. J. G. González, and I. M. Oliver. Unsupervised domain adaptation for the segmentation of breast tissue in mammography images, 2021. Computer Methods and Programs in Biomedicine, 211, 106368.

[98] 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.

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

[100] 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.

[101] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018.

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

[103] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[104] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A Prototype-Oriented Framework for Unsupervised Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17194–17208, 2021.

[105] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, page 1195–1204, 2017.

[106] 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.

[107] 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.

[108] Arina Varlamova, Valery Belotsky, Grigory Novikov, Anton Konushin, and Evgeny Sidorov. Features fusion for dual-view mammography mass detection. *arXiv preprint*, 2024.

[109] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

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

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

[117] Y. Wang and Y. Yao. Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. *Scientific Reports*, 12(1):14720, 2022.

[118] Yu Wang and Yudong Yao. Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement, December 2022. Publisher Copyright: © 2022, The Author(s).

[119] 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.

[120] 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.

[121] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.

[122] 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.

[123] 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.

[124] 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, 05 2021.

[125] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020.

[126] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H^2FA R-CNN: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, 2022.

[127] 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.

[128] Ke Yan, Youbao Tang, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M. Summers. Mulan: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation, 2019.

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

[130] 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.

[131] 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.

[132] Yuxiang Yang, Xinyi Zeng, Pinxian Zeng, Binyu Yan, Xi Wu, Jiliu Zhou, and Yan Wang. Btmuda: A bi-level multi-source unsupervised domain adaptation framework for breast cancer diagnosis, 2024.

[133] 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.

[134] 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.

[135] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2020.

[136] 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.

[137] 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 *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 629–645, Berlin, Heidelberg, 2022. Springer-Verlag.

[138] 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.

[139] 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.

[140] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. *AAAI*, 36(3):3252–3261, 2022.

[141] 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.

[142] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[143] 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.

[144] 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, June 2022.

[145] Sicheng Zhao, Huizai Yao, Chuang Lin, Yue Gao, and Guiguang Ding. Multi-source-free domain adaptive object detection. *Int. J. Comput. Vision*, 132(12):5950–5982, July 2024.

[146] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1):166–176, November 2024.

[147] 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.

[148] 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18993–19003, 2023.

[149] 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.

[150] 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.

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

[152] 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.

[153] 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.

[154] 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.