Domain Adaptation using Self-Training with Mixup for One-Stage Object Detection

Jitender Maurya, Keyur R. Ranipa
Toshiba Software India Pvt. Ltd.
{jitender.maurya, keyur.ranipa}@toshiba-tsip.com

Osamu Yamaguchi, Tomoyuki Shibata, Daisuke Kobayashi
Corporate Research and Development Center, Toshiba Corporation
{osamu1.yamaguchi, tomoyuki1.shibata,daisuke32.kobayash}@toshiba.co.jp

Abstract

In this paper, we present an end-to-end domain adaptation technique that utilizes both feature distribution alignment and Self-Training effectively for object detection. One set of methods for domain adaptation relies on feature distribution alignment and adapts models on an unlabeled target domain by learning domain invariant representations through adversarial loss. Although this approach is effective, it may not be adequate or even have an adverse effect when domain shifts are large and inconsistent. Another set of methods utilizes Self-Training which relies on pseudo labels to approximate the target domain distribution directly. However, it can also have a negative impact on the model performance due to erroneous pseudo labels. To overcome these two issues, we propose to generate reliable pseudo labels through feature distribution alignment and data distillation. Further, to minimize the adverse effect of incorrect pseudo labels during Self-Training we employ interpolation-based consistency regularization called mixup. While distribution alignment helps in generating more accurate pseudo labels, mixup regularization of Self-Training reduces the adverse effect of less accurate pseudo labels. Both approaches supplement each other and achieve effective adaptation on the target domain which we demonstrate through extensive experiments on one-stage object detector. Experiment results show that our approach achieves a significant performance improvement on multiple benchmark datasets.

1. Introduction

Object detection is a significant task in computer vision. With recent advances in deep learning and the availability of large-scale datasets, the deep Convolutional Neural Networks (CNNs) based object detectors [17, 8, 21] have significantly improved the performance of the object detection task on benchmark datasets [6, 16]. However, it is challenging to directly utilize an object detector trained on these benchmark datasets (source dataset) in an application where the dataset (target dataset) has a significant domain gap. This problem can be addressed with augmenting the source dataset with annotated target dataset. However, obtaining object annotations for each new domain dataset is time consuming, costly, and requires significant human efforts.

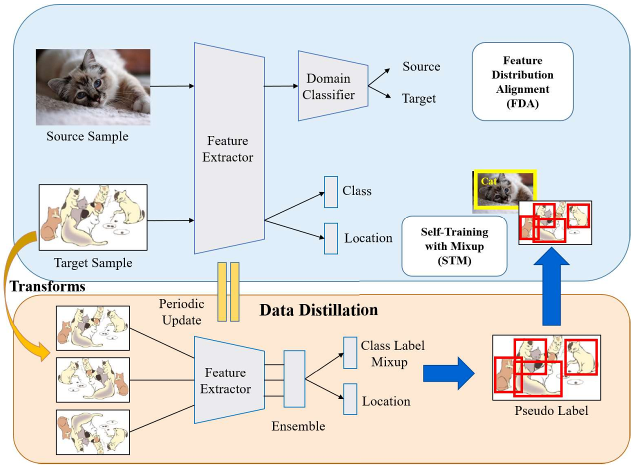

Figure 1. Illustration of the proposed method. **Top:** Applying Feature Distribution Alignment to reduce domain shifts and Self-Training with mixup. **Bottom:** Applying data distillation to generate pseudo labels and then using them during Self-Training to compute detection loss on target domain samples.

Unsupervised Domain Adaptation (UDA) is an approach to address the issue of domain shifts between label-rich source domain dataset and label-scarce target domain. Various methods for UDA have been proposed, of which most aim to reduce domain shifts by aligning source and target domain feature distributions [7, 19, 26]. While these methods focus on the classification task, UDA methods for object detection have recently gained much attention [4, 25, 14, 37, 27, 33, 2, 9, 32, 36]. These state-of-the-art UDA methods for domain adaptive object detection attempt to align feature distribution of source and target domain by introducing adversarial loss. In this paper, training a model with an adversarial loss is denoted as Feature Distribution Alignment (FDA) step. Although these methods achieve a considerable performance improvement in the target domain over baseline object detectors, it remains challenging to further boost the performance when the domain gap is large. Especially, in the case of object detection when instance level, intra-domain variations are higher, feature distribution alignment can be ineffective or might even exhibit an adverse performance.

Another approach to domain adaptive object detection is Self-Training [23, 14, 22]. Self-Training utilizes a model trained on label-rich source domain dataset to generate pseudo labels for target domain. These labels are then used to train a model on target domain to approximate real data distribution. However, traditional approaches of Self-Training can have a negative effect on the model performance for multiple reasons. First, due to domain gap, generated pseudo labels are erroneous. Second, these incorrect pseudo labels can reinforce themselves during the subsequent iterations of Self-Training and further amplify the error, resulting in the negative performance. Finally, these labels are generated from model's own predictions and may not have any meaningful information for improving performance on the target domain. These issues lead to rapid degradation in the performance of the model when Self-Training is applied for a longer period.

In the proposed method we address these two issues. First, as shown in Figures 1 and 2, we utilize FDA as a supplement step which helps to generate reliable pseudo labels for the Self-Training. Due to FDA, model learns domain invariant feature representations which helps in generating more accurate pseudo labels of target domain samples. In FDA, we apply feature distribution alignment at multiple feature levels of an object detector. Then we use model trained with FDA to generate pseudo labels that aid the Self-Training on the target domain. To further improve the accuracy of pseudo labels, we utilize data distillation. Data distillation by ensembling predictions of the same model on multiple transformations of the input is a common strategy to boost the performance of the model at a test time [20]. In the proposed Periodic Data Distillation (PDD), pseudo labels are generated by ensembling the predictions of the model on multiple transformations of the target domain images. These labels are periodically updated during the Self-Training as the model with FDA gets better at predictions. The combination of FDA, that reduces the domain gap between the source and target domain and PDD, helps in generating more accurate pseudo labels.

Further, to address the issue of less accurate pseudo labels, we propose a consistency regularization of Self-Training with mixup (STM) [35, 30]. Mixup is a regularization technique using interpolation-based data augmentation which provides robustness against incorrect pseudo labels. We increase the strength of mixup interpolation to generate samples further from the real training samples to make memorization of corrupt or random labels difficult for the model. In contrast, when pseudo labels are correct and consistent it is easier to learn from such samples. Further, it also enforces consistent predictions on samples which lie on decision boundary [31]. Therefore, in the proposed method we employ mixup regularization to enforce consistent predictions and reduce memorization of incorrect pseudo labels during Self-Training.

We demonstrate the effectiveness of the proposed method on one-stage object detector SSD (Single Shot MultiBox Detector) [17]. Simplicity and a higher frame rate make one-stage object detectors more suitable for real-world applications. Further, domain adaptation methods based on one-stage detectors are limited. Therefore, we apply the proposed method to one-stage object detector. The main contributions of this paper are summarized as follows:

- We propose a method to generate accurate pseudo labels through Feature Distribution Alignment and Periodic Data Distillation.
- We propose an end-to-end, Self-Training method with interpolation-based consistency regularization called mixup and confirm its validity in the context of domain adaptation.
- We validate the proposed method on a one-stage detector. We conduct extensive experiments and achieve a comparable performance to state-of-the-art methods on the standard domain adaptation benchmarking datasets such as Clipart1k, Watercolor2k, Comic2k and challenging real-world scenario of domain adaptation for the task of pedestrian detection.

2. Related Work

Domain Adaptation addresses the issue of domain shifts between the source and target domain. State-of-the-art domain adaptation techniques for object detectors can broadly be classified in three categories. First are domain transfer techniques where source domain images are first translated into target domain and then the detector is fine-tuned on both source and target domain dataset. Domain transfer techniques were utilized in [10, 22] to perform two-step

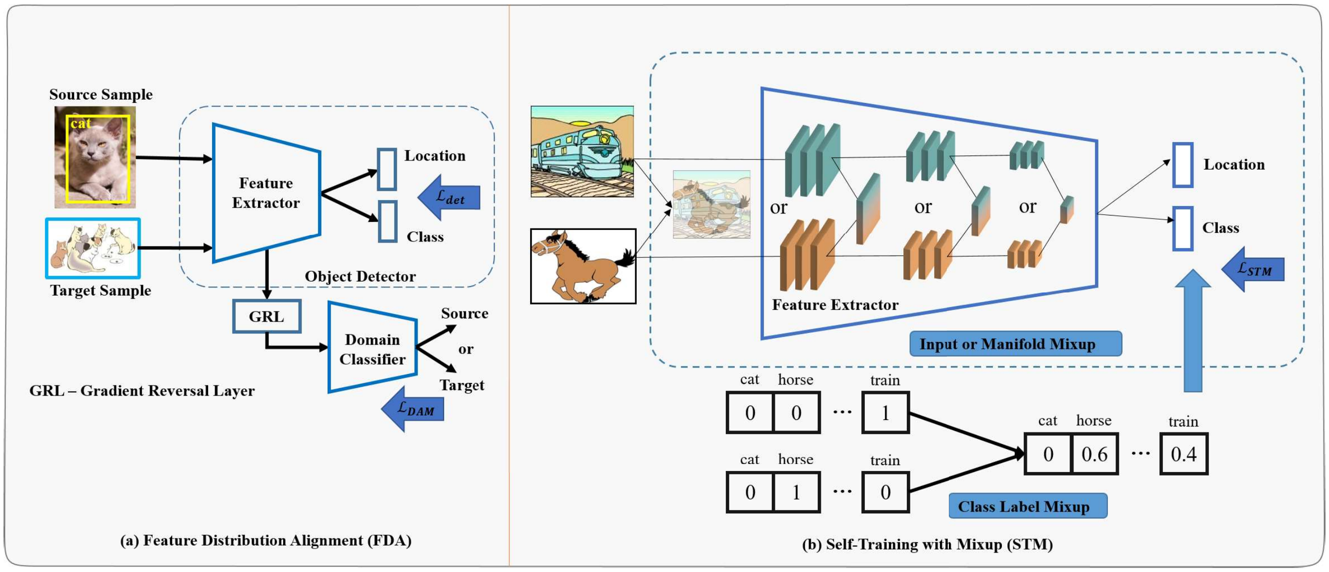

Figure 2. Illustrating components of the proposed method. **(a)** Feature Distribution Alignment (FDA) using domain classifier and adversarial training. For simplicity, only one domain classifier is illustrated. **(b)** Self-Training with mixup regularization (STM) **Top:** Mixup is performed only on a single eligible layer selected (among input, or hidden representations) randomly at each iteration. **Bottom:** Demonstrating object class label mixup. For clarity, images with one object are illustrated. If images contain more objects, mixing of each object class label is performed. Best viewed in color.

progressive domain adaptation on one-stage object detector SSD[17]. They adapt source images to target domain using style transfer [38], afterward features are aligned to reduce domain gap. The second set of methods, that aligns distribution of source and target domain using an adversarial loss, has been very effective in classification and object detection tasks [14, 7, 4, 25, 18, 26, 33]. For example, weak global alignment method proposed in [25] focuses on aligning features of hard-to-classify samples in target domain and Background Score Regularization (BSR) proposed in [14] uses adversarial loss to train a feature extractor to extract discriminative features of target domain. Implicit Instance-Invariant Network (I^{3}Net) [3] is specifically tailored for the one-stage detector, which considers intra-class, intra-domain variations and object patterns to learn instance invariant features.

The third technique is based on Self-Training, which uses pseudo labels to guide training over unlabeled target-domain data. Self-Training has been used along with both, a domain transfer technique [22, 10, 34] and adversarial training technique [14]. In general, the approach in these methods is to reduce domain gap for the purpose of generating reliable pseudo labels and then using them to self-train the object detector on target domain dataset.

Mixup is an interpolation-based regularization method that has been used for both supervised [30, 35] and semi-supervised learning settings [31, 1]. Input mixup [35] performs an interpolation of input space, while manifold mixup [30] performs an interpolation of hidden state representation to improve robustness against adversarial samples in supervised learning settings. In semi-supervised learning of image classification task, it is used [31] to enforce consistent and low-entropy predictions [1] on interpolation of unlabeled samples. Further, for the task of object detection, [12] shows the effectiveness of interpolation-based regularization in semi-supervised learning settings, where primary objective is to improve performance of the detector on unlabeled samples of the source domain dataset. In contrast, the proposed method focuses on improving the performance of the object detector on the target domain dataset which does not have any labeled samples.

The proposed method primarily employs two techniques for building a domain adaptive object detector, Feature Distribution Alignment (FDA) and Self-Training (ST). We utilize FDA to reduce domain shifts and thus generate more accurate pseudo labels of target domain samples. These pseudo labels are then used in Self-Training of object detector on the target domain. Self-Training can be highly unstable and can have detrimental effect on the performance of the model in presence of incorrect pseudo labels. Therefore, we apply mixup based consistency regularization to reduce the adverse effect of incorrect pseudo labels.

3. Proposed Method

In Unsupervised Domain Adaptation setting of object detectors, we are given a set of images X_s from the source domain with labels Y_s and unlabeled set of images X_t from the target domain. (x_s, y_s) is a sample drawn from X_s, where x_s is an image and $y_s = \{(c_i, l_i)_s\}_{i=1}^{m}$ are corresponding m object labels, where c_i is the class label and l_i is the location of the i^{th} object within image. Further, x_t is a sample drawn from X_t and corresponding object labels are not available. $P(X_s)$ and $P(X_t)$ denote source and target domain distributions respectively and $P(X_s) \neq P(X_t)$. The illustration of components of the end-to-end proposed method is shown in Figure 2.

3.1. Feature Distribution Alignment

The primary objective of the Feature Distribution Alignment is to reduce the large domain distribution gap between $P(X_s)$ and $P(X_t)$. To achieve this objective, various methods have been proposed [7, 19] that align feature distribution by minimizing the distance between two domains. In general, this is done through attaching domain classifiers to feature extractor at various feature levels [25]. Then, the feature extractor and domain classifiers are trained in an adversarial way. This approach has been applied for various vision applications such as image classification and object detection. Recently, Saito *et al.* [25] proposed a *Strong* and *Weak* domain adaptation for the two-stage detector Faster RCNN [21]. They apply a *Weak* alignment to global features to alleviate the negative effect of feature alignment when domain shifts are large. For *Weak* alignment, they utilize Focal Loss (FL) [15] instead of cross-entropy loss as an objective for the domain classifier of global level features. For local features, such as texture and color, at lower layers of feature extractor, they apply *Strong* alignment by employing least-squares loss as an objective for the domain classifier. In the proposed method, FDA step is based on this idea but for one-stage object detector Figure 2 (a). We employ *Weak* alignment at global level features of the feature extractor. However, for the *Strong* alignment of local level features, we employ two domain classifiers at two different lower layers of the feature extractor. This helps in achieving alignment features not just at local level, but also of features at intermediate level which might constitute object features. By reducing the domain shifts, FDA helps in generating more reliable pseudo labels for the target domain dataset as compared to the model that is trained only on the source domain dataset. This is the primary motivation behind adding this module in the proposed method. The adversarial loss of FDA step is summarized as follows:

$$\mathcal{L}_{FDA} = \mathcal{L}_{global} + \mathcal{L}_{loc_1} + \mathcal{L}_{loc_2} \tag{1}$$

where, $\mathcal{L}_{global}$ is same as proposed in [25]. As we have two local classifiers at different layers, losses corresponding to them are denoted as $\mathcal{L}_{loc_1}$ and $\mathcal{L}_{loc_2}$ and they are also as proposed in [25].

3.2. Pseudo Labels using Periodic Data Distillation

As FDA reduces a distribution gap, model gets better at predicting correct labels for the target domain images. However, a single prediction from the model may not always be accurate. Moreover, misclassified samples with high-confidence score may end up being pseudo labels. Such pseudo labels can have a negative impact on the performance as they have a capacity to amplify the error during Self-Training. Accuracy of pseudo labels is of utmost important for effective and stable Self-Training. Therefore, we propose a Periodic Data Distillation method where a multi-transform inference [20] is applied on the model to generate high-quality pseudo labels. Data distillation using multi-transform inference has been proven effective on variety of vision applications[11, 13, 20] including object detection. In data distillation, a trained model is applied on the multiple transformations of unlabeled input image to get predictions. The predicted bounding boxes for each object from all the transformed images are ensembled to generate a pseudo label. This process will help to have correct predictions as it is more likely to have bounding boxes in a common region in most predictions.

Given a state of the model F_{θ}^{T-1} at epoch $T-1$ with model parameters θ, and unlabeled target image sample x_t with its set of k geometric transformations $\{x_t^1, x_t^2, \ldots, x_t^k\}$, where superscript 1 represents original image without any transformations, then the pseudo label is produced as follows:

$$y_t = \mathit{ensemble}\{F_{\theta}^{T-1}(x_t^1), F_{\theta}^{T-1}(x_t^2) ... F_{\theta}^{T-1}(x_t^k)\} \tag{2}$$

Here, $y_t = \{(c_i, l_i)_t\}_{i=1}^{m}$ will be used as a pseudo label for x_t during Self-Training of the model state F_{θ}^{T} at epoch T.

To generate pseudo labels, we first apply multi-transform inference on the image x_t. For each transformed image there are N_c object predictions of class c. If the location of the object is represented by bounding box coordinates, then $B_c = \{B_i\}_{i=1}^{N_c}$ is a set of all such bounding boxes in the image. In the $\mathit{ensemble}$ operation, we merge k predictions $\tilde{B}_c = \bigcup_{j=1}^{k} B_c = \{B_i\}_{i=1}^{N_{ck}}$ of multi-transform inference. Because of the multi-transform inference, the generated boxes will be highly concentrated around the actual objects of interest. We exploit this aspect in the $\mathit{ensemble}$ operation and then, the final bounding box coordinates $B_{i,c}^{'}$ is an average

$$B_{i,c}^{'} = \frac{B_{i,c} + \sum_{\mathcal{N}(B_{i,c})} B_{j,c}}{1 + \mathit{count}(\mathcal{N}(B_{i,c}))} \tag{3}$$

coordinates of each bounding box $B_{j,c} \in \mathcal{N}(B_{i,c})$ (where $\mathcal{N}(B_{i,c})$ denotes the set of boxes in $\tilde{B}_c$ that overlap with $B_{i,c}$ by more than 0.5 on IoU metric).

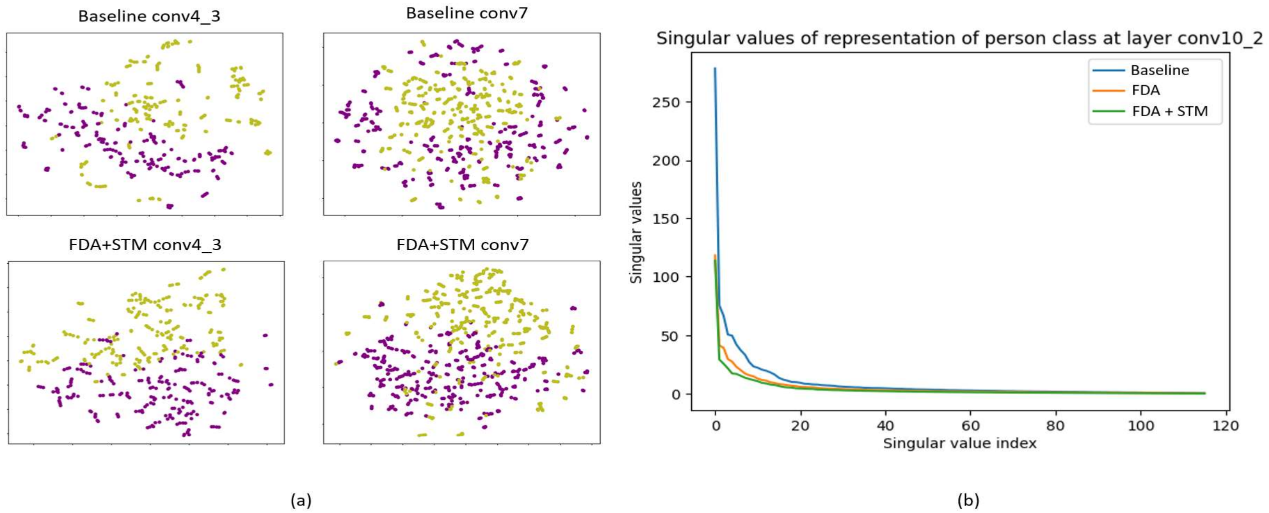

Figure 3. A visualization on Clipart1k feature representations **(a) Top:** Feature representations of Person (purple-colored dots) and Chair class (olive-colored dots) learned by Baseline model at *conv4_3* and *conv7* layer. **Bottom:** Feature representation learned by FDA+STM for same classes. It can be observed that Self-Training has improved discriminative capabilities of the model on target domain. (**b**) Singular Value Decomposition of class specific representations at *conv10_2* layer. With FDA+STM, it can be observed that it reduces the number of directions with significant variance. Best viewed in color.

3.3. Consistency Regularization with Mixup

Even though FDA and PDD produce more accurate pseudo labels, the problem of incorrect pseudo labels remains. Just few of these labels can have a substantial adverse effect on model performance as they can reinforce themselves during the subsequent iterations of Self-Training. To address this issue, we propose a consistency regularization of Self-Training with mixup. We aim to achieve two objectives through mixup regularization. First is to reduce memorization of incorrect pseudo labels for which we increase the strength of the mixup interpolation. Second objective is to enforce consistent predictions on boundary samples (which are most likely to be misclassified) in the target domain and improve generalization capability of the model which is achieved implicitly through mixup of random unlabeled target samples and their corresponding pseudo labels. Perturbation by interpolation have been proven effective for consistency regularization in Semi-Supervised Learning settings [31]. In case of Self-Training, mixup operation functions as a regularizer which enforces consistent predictions on unlabeled samples even after it has been augmented through interpolation.

Mixup operation, as described in [35, 31], for a pair of target samples (x_t^i, y_t^i) and (x_t^j, y_t^j) is as follows:

$$\begin{aligned}\hat{x}_t &= \lambda x_t^i + (1-\lambda)x_t^j \\ \hat{y}_t &= \lambda y_t^i + (1-\lambda)y_t^j,\end{aligned} \tag{4}$$

where λ is a mixup factor which, following [35], is drawn from the $Beta(\alpha, \alpha)$ distribution. As discussed before, to minimize the effect of incorrect pseudo labels we choose higher value of α to increase the strength of the interpolation. Value of α is a hyper-parameter, which we set to 16.0 in all our experiments as it gives the best results. Here, it should be noted that the pseudo label $y_t = \{(c_i, l_i)\}_{i=1}^{m}$ of x_t have m class labels corresponding to m objects. Therefore, unlike mixup operations in classification task [35, 30] where mixup is applied on image labels, we apply mixup operation on object labels. Further, mixup is applied only on the class labels of the objects and not on location labels.

3.4. Self-Training with Mixup (STM)

For the unlabeled target dataset, Self-Training enables learning representations directly from the dataset with the help of pseudo labels. During Self-Training we apply random mixup in which mixup factor λ is drawn randomly from $Beta(\alpha, \alpha)$ distribution for each iteration. Further, instead of applying mixup only on a pair of target inputs (x_t^i, x_t^j), following [30], we apply mixup on both inputs and their hidden representations (2 (b)). These hidden representations are also selected randomly following [30]. During Self-Training pseudo labels are periodically updated at every epoch of the target dataset as the model gets better due to the effect of FDA and Self-Training. Finally, with pseudo label mixup the objective of the Self-Training is defined as follows:

$$\mathcal{L}_{STM} = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_t^i \log(\tilde{y}_t^i) - (1-\hat{y}_t^i)\log(1-\tilde{y}_t^i) \tag{5}$$

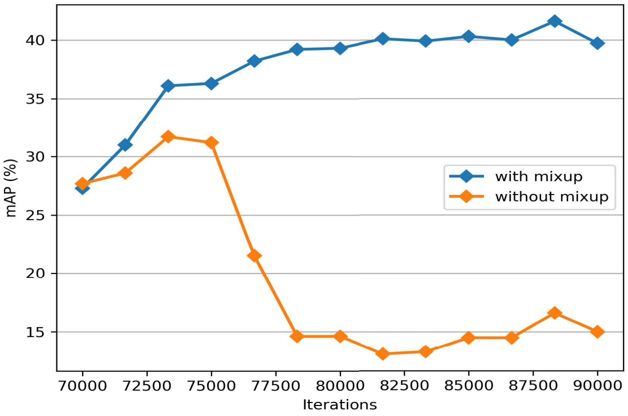

Figure 4. Showing trends of mAP at various iterations during Self-Training period. Self-Training without mixup regularization (orange) has adverse effect on mAP when applied for a long duration. The proposed mixup regularization (blue) stabilizes the Self-Training and mAP improves steadily.

where $\tilde{y}_t^i = p(y|\hat{x}_t^i)$ is a probability distribution of the class prediction produced by the model on the input $\hat{x}_t^i$ and N is the number of samples in a mini-batch. Please see our supplemental material for the discussion on using localization loss to regress bounding boxes during Self-Training.

3.5. Model Training

The training of the object detector happens in two phases. In the first phase, the model is trained with FDA to reduce the distribution shifts between the source and the target datasets. Once the model has been stabilized after training it for a certain number of iterations, in the second phase we enable the Self-Training. In all our experiments, Self-Training runs for 10k to 20k iterations. The overall objective of the training is as follows:

$$\mathcal{L} = L_{det} + \lambda_{FDA} L_{FDA} + \lambda_{STM} L_{STM} \tag{6}$$

where λ_{STM} is set to 1 only during the Self-Training otherwise set to 0, L_{det} is SSD [17] objective on the source dataset, L_{FDA} is FDA loss as described in Equation 1, λ_{FDA} is a domain adaptation factor which gradually changes from 0 to 1 during training, as described in [7].

4. Experiments

In this section we describe experiments and results of the proposed method on three domain adaptation benchmarks for object detection. We also present the effectiveness of the proposed method in real-world domain adaptation scenario for the application of pedestrian detection.

4.1. Implementation

For all experiments, we use one-stage object detector SSD as a baseline detector. The baseline detector employs a VGG16 [28] as a feature extractor pre-trained on the ImageNet [24]. The input images are resized to 300x300. It is trained with Stochastic Gradient Descent (SGD) optimizer, with a momentum of 0.9 and a weight decay of 0.0005. The base learning rate is set to 0.01 with the batch size of 8. For all experiments we train the model for 120k iterations.

For training the model with FDA, global level features are extracted from the first detection head of SSD which is at *conv4_3* layer in case of VGG16. These features are fed to the global domain discriminator. The local level features are extracted from shallow layers *conv3_2* and *conv3_3* of VGG16 and fed to local domain discriminators. To apply the adversarial training strategy, Gradient Reversal Layer (GRL) is added before each of the domain discriminators.

Self-Training is enabled between 70k-90k iterations of training. During Self-Training pseudo labels are updated at every epoch of the target dataset. In data distillation, we set $k = 5$ to ensemble predictions from original image and 4 geometric transformations on the same image. These transformations are horizontal-flip, vertical-flip, down-scale, and up-scale.

4.2. Datasets

For adaptation, PASCAL VOC [6] has been used as the source domain dataset. This dataset has object class annotations of 20 classes and their bounding boxes. We use PASCAL VOC 2007 and 2012 training and validation splits for training. The target domain consists of Clipart1k [10], Comic2k [10] and Watercolor2k [10] datasets. Clipart1k contains 1k images in total, which have the same 20 categories as PASCAL VOC. 500 images were used for training and 500 for testing. Watercolor2k and Comic2k contains 6 categories in common with PASCAL and 2k images in total. 1k training images were utilized during training and 1k images are used as test images.

To test the effectiveness of our proposed method in real-world scenario, we also show results of domain adaptation from COCO [16] to Caltech [5] for the task of pedestrian detection. Caltech dataset is a pedestrian detection benchmark dataset having 16k train samples and 2.24k test samples. For the task of pedestrian detection, detectors are trained only on a person class therefore we use images having person class from the COCO, resulting in around 45k train samples. For more details on Caltech dataset, please see our supplemental material.

4.3. Results

We employ COCO evaluation [16] protocol to measure the mAP of the model, where we set the confidence threshold to 0.05 and the IoU threshold to 0.5. We compare the proposed method with the Source only model [17] which is trained only on the source dataset, DANN [7], Background Score Regularization and Weak Self-Training [14] (denoted

Table 1. Results on the Clipart1k test dataset and comparison of various methods in terms of mAP (%). To adapt SSD [17] on Clipart1k, PASCAL VOC is used as a source dataset.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Source Only[17]	27.3	60.4	17.5	16.0	14.5	43.7	32.0	10.2	38.6	15.3	24.5	16.0	18.4	49.5	30.7	30.0	2.3	23.0	35.1	29.9	26.7
DANN[7]	24.1	52.6	27.5	18.5	20.3	59.3	37.4	3.8	35.1	32.6	23.9	13.8	22.5	50.9	49.9	36.3	11.6	31.3	48.0	35.8	31.8
BSR+WST[14]	28.0	64.5	23.9	19.0	21.9	64.3	43.5	**16.4**	42.2	25.9	30.5	7.9	25.5	67.6	54.5	36.4	10.3	31.2	**57.4**	43.5	35.7
I^{3}Net[3]	**30.0**	67.0	**32.5**	21.8	29.2	62.5	41.3	11.6	37.1	39.4	27.4	19.3	25.0	67.4	55.2	**42.9**	**19.5**	36.2	50.7	39.3	37.8
DBGL[2]	23.2	65.5	30.1	18.3	24.6	**67.6**	43.9	15.1	38.7	36.4	31.3	20.2	25.0	**74.3**	55.1	38.2	12.5	**41.0**	49.1	**43.9**	37.7
STM	28.2	65.2	14.8	17.9	3.1	32.4	39.1	5.8	45.4	7.8	35.5	15.8	23.8	44.5	55.2	36.4	5.0	23.6	35.1	37.7	28.6
FDA	28.5	63.1	24.4	22.0	18.8	35.1	38.0	2.5	**47.7**	43.5	23.7	8.4	31.3	49.4	50.3	32.9	12.7	32.9	35.0	39.4	32.0
FDA+STM	26.9	**76.7**	30.7	**34.8**	**32.9**	52.0	**49.5**	6.3	46.8	**55.4**	**45.1**	**21.5**	**41.0**	66.3	**70.0**	42.0	11.7	26.4	51.8	43.8	**41.6**

Table 2. Results on Comic2k test dataset in terms of mAP(%).

Method	bike	bird	car	cat	dog	person	mAP
Source Only[17]	43.3	9.4	23.6	9.8	10.9	34.2	21.9
DANN [7]	33.3	11.3	19.7	13.4	19.6	37.4	22.5
BSR+WST [14]	50.6	13.6	31.0	7.5	16.4	41.4	26.8
I^{3}Net [3]	47.5	**19.9**	**33.2**	11.4	19.4	49.1	30.1
DBGL[2]	45.4	15.9	24.8	11.5	29.4	55.1	30.4
STM	42.2	11.2	15.5	17.5	8.7	40.1	22.5
FDA	51.1	15.9	12.1	24.8	25.3	50.4	29.9
FDA+STM	**51.8**	16.4	13.1	**34.6**	**32.6**	**63.8**	**35.4**

Table 3. Results on Watercolor2k test dataset in terms of mAP(%).

Method	bike	bird	car	cat	dog	person	mAP
Source Only[17]	77.5	46.1	44.6	30.0	26.0	58.6	47.1
DANN [7]	73.4	41.0	32.4	28.6	22.1	51.4	41.5
BSR+WST [14]	75.6	45.8	**49.3**	34.1	30.3	64.1	49.9
I^{3}Net [3]	81.1	49.3	46.2	35.0	31.9	65.7	51.5
DBGL[2]	84.0	46.7	45.5	36.2	35.7	63.7	52.0
STM	91.8	50.6	34.4	47.1	37.3	72.7	55.7
FDA	83.1	46.6	29.5	37.0	32.6	67.4	49.4
FDA+STM	**99.3**	**52.7**	44.5	**49.1**	**38.8**	**73.3**	**59.6**

Table 4. Results on Caltech dataset for the task of pedestrian detection in terms of mAP (%).

Method	Person
Source Only[17]	27.1
DANN [7]	26.8
STM	**32.7**
FDA	28.6
FDA+STM	32.4

as BSR+WST), I^{3}Net [3] and DBGL[2]. BSR+WST, I^{3}Net and DBGL are implemented on SSD and have evaluation protocol identical to ours. The results of DANN and BSR+WST are cited from [14], results of I^{3}Net are cited from the original paper [3] and results of DBGL are cited from the original paper[2]. We also demonstrate the effectiveness of individual components of the proposed method which are Feature Distribution Alignment (FDA) and Self-Training with mixup regularization (STM). They are denoted as FDA and STM in the result tables. In results, FDA+STM indicates both components are used in training the model.

Clipart1k. As shown in Table 1 the mAP of the proposed FDA+STM is 41.6%, which is an improvement of +14.9% over the Source only model. When only FDA is applied the mAP is 32.0% and applying STM improves the performance of the model by +9.6%. This is a significant improvement and shows the effectiveness of the proposed Self-Training method. The proposed method also outperforms BSR+WST [14], I^{3}Net [3] and DBGL[2] by considerable margin. Further, it can be observed that even when STM is applied without FDA, mAP is improved. This shows that mixup is effective in regularizing the Self-Training and it reduces the adverse effect of incorrect pseudo labels.

Comic2k. As shown in Table 2, on Comic2k dataset mAP of FDA+STM is 35.4%. In comparison to BSR+WST [14], I^{3}Net [3] and DBGL[2], the proposed method outperforms them by a significant margin of +8.6%, +5.3% and +5% respectively. Overall improvement in mAP is 13.5% over the Source only model.

Watercolor2k. Results on Watercolor2k dataset are shown in Table 3. When only FDA is applied, there is a marginal improvement in mAP. After employing STM along with FDA, the mAP is further increased by +10.2%. We also observe that when only STM is applied, the mAP is improved by +8.6% over the Source only model. It suggests that when FDA is not very effective due to high intra-domain variance, STM is helpful to boost the performance. The proposed method significantly outperforms state-of-the-art BSR+WST [14], I^{3}Net [3] and DBGL[2] method and achieve the best mAP.

Caltech Pedestrian Benchmark. In this case, we examine the effectiveness of the proposed method in real-world domain adaptation scenario. We use COCO [16] dataset as a source domain and train SSD [17] on it to detect objects of only a person class. The mAP obtained is 61.6%. When the test dataset of Caltech is evaluated with the Source only model, it gives a mAP of 27.1%. COCO is a large-scale dataset in which images are collected from various sources and have high intra-domain variance. Therefore, it can be observed in Table 4, that the adaptation by FDA is not helpful. Employing STM along with FDA is very effective and improves a mAP by +5.3% to 32.4%. These results show that the STM is effective in challenging real-world scenarios too.

Table 5. mAP (%) of target test sets for different mixup strength.

Method	Clipart1k	Comic2k	Watercolor2k
Source Only[17]	26.7	21.9	47.1
FDA+STM ($\alpha = 2.0$)	40.0	32.4	54.0
FDA+STM ($\alpha = 4.0$)	39.9	31.6	57.1
FDA+STM ($\alpha = 8.0$)	40.2	**35.4**	57.4
FDA+STM ($\alpha = 16.0$)	**41.6**	**35.4**	**59.6**

4.4. Discussion and Ablation Studies

In this section, we present the analysis of the proposed method. We conducted the ablation studies on Self-Training, analyze the hyper-parameter sensitivity of α which controls the mixup ratio, and study the effect of disabling data distillation during Self-Training.

Effect of mixup regularization during Self-Training. We investigate the effectiveness of mixup regularization during Self-Training. While training SSD with FDA and Self-Training on adaptation from PASCAL VOC to Clipart1k, we measure the mAP of the model on Clipart1k at iterations during Self-Training period. As we can see in Figure 4, with mixup regularization, Self-Training is highly stable without adverse impact on the performance even when it runs for very long duration. In contrast, without mixup regularization during Self-Training, the performance improves momentarily for few iterations and then it rapidly degrades due to reinforcement of incorrect pseudo labels. This shows mixup is useful in minimizing the impact of incorrect pseudo labels.

Effect of mixup parameter α. The hyper-parameter α controls the mixup ratio drawn from $\lambda \sim Beta(\alpha, \alpha)$ distribution. Increasing the value of α will increase the strength of interpolation and will generate the interpolated samples further from the real samples. In effect, if labels of such samples are random or corrupt then it becomes difficult for the network to learn such associations. Thus, minimizing the effect of incorrect pseudo labels. We trained the model with different values of mixup α and results are shown in Table 5 on Clipart1k, Comic2k and Watercolor2k datasets.

Effect of Periodic Data Distillation. To analyze the effectiveness of PDD, we disable it and train the model with FDA and STM on Clipart1k, Comic2k and Watercolor2k datasets. In this experiment, we do not use multi-transform inference on the model to generate pseudo labels and predictions only from the original target samples are used as pseudo labels. As shown in Figure 5, we observe that disabling the data distillation reduces the performance of the model on all three benchmarks. It shows that Periodic Data Distillation is beneficial in generating more accurate pseudo labels and helps in achieving the optimal performance.

Feature Visualization. Figure 3 (a) presents the visualization of t-SNE [29] feature representations of two classes learned by Baseline and FDA+STM model. Baseline model has lower discriminative ability on target domain classes. Feature representations at both *conv4_3* and *conv7* layers have overlapping space due to which detector has low classification accuracy. On the other hand, after Self-Training with mixup regularization, both classes have a separate feature space which improves discriminative capability of the model on target domain.

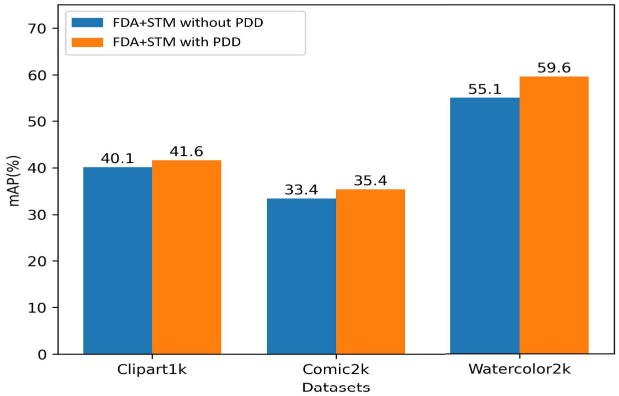

Figure 5. Comparison of FDA+STM mAP(%) in two different settings. FDA+STM with PDD (orange bars) outperforms FDA+STM without PDD (blue bars) on all three benchmarks.

Reducing directions of variance. In Figure 3 (b) we analyze *flattening* property of mixup regularization. Authors of [35, 30] show that mixup regularization leads to *flattening*: a reduction in the number of directions with variability [30]. Reducing directions of variability improves object classification ability of the detector. To investigate *flattening*, we compute Singular Value Decomposition of class specific representation at layer *conv10_2*. As compared to Baseline, in FDA+STM, we observe that large singular values are concentrated at lower indices which means that class specific representation has lower variance.

5. Conclusion

In this paper, we proposed a domain adaptation method using FDA and Self-Training for one-stage object detector. Our method consists of two major components. First is FDA, which utilizes adversarial loss to reduce the domain shifts between source and target domain. As domain shifts reduce, the model gets better at predictions on target domain samples. Taking advantage of this property, we use PDD to generate more accurate pseudo labels of the target domain. In the second component, STM, we use pseudo labels to guide the Self-Training of an object detector and regularize Self-Training with mixup. These two enable a stable and effective Self-Training and achieve state-of-the-art performance on domain adaptation benchmarks. Further, we validate the application of the proposed method in real-world scenario for the task of pedestrian detection where it shows a significant improvement in the performance.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[2] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021.

[3] 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 (CVPR)*, pages 12576–12585, June 2021.

[4] 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.

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 17(1):2096–2030, Jan. 2016.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1440–1448, USA, 2015. IEEE Computer Society.

[9] Luwei Hou, Yu Zhang, Kui Fu, and Jia Li. Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9929–9938, 2021.

[10] 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 (CVPR)*, June 2018.

[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[12] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. *CoRR*, abs/2006.02158, 2020.

[13] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019.

[14] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 97–105. JMLR.org, 2015.

[19] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 136–144. Curran Associates Inc., 2016.

[20] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omnisupervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[22] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency, 2019.

[23] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[25] 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 (CVPR)*, June 2019.

[26] 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 (CVPR)*, June 2018.

[27] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses, 2019.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[30] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[31] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3635–3641. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[32] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, Yangyang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9603–9612, 2021.

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2022.

[34] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, and Xiang Chen. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning, 2021.

[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, Apr. 2018.

[36] 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*, pages 14217–14226, 2022.

[37] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.