Unsupervised Domain Adaptive Object Detection using Forward-Backward Cyclic Adaptation

Siqi Yang[1], Lin Wu[2], Arnold Wiliem[1], and Brian C. Lovell[1]

[1] The University of Queensland, Brisbane, Australia
[2] Hefei University of Technology, Hefei, China
siqi.yang@uq.net.au, xiaoxian.wu9188@gmail.com, arnold.wiliem@ieee.org, lovell@itee.uq.edu.au

Abstract. We present a novel approach to perform the unsupervised domain adaptation for object detection through forward-backward cyclic (FBC) training. Recent adversarial training based domain adaptation methods have shown their effectiveness on minimizing domain discrepancy via marginal feature distributions alignment. However, aligning the marginal feature distributions does not guarantee the alignment of class conditional distributions. This limitation is more evident when adapting object detectors as the domain discrepancy is larger compared to the image classification task, *e.g.,* various number of objects exist in one image and the majority of content in an image is the background. This motivates us to learn domain-invariance for category-level semantics via gradient alignment for instance-level adaptation. Intuitively, if the gradients of two domains point in similar directions, then the learning of one domain can improve that of another domain. We propose Forward-Backward Cyclic Adaptation to achieve gradient alignment, which iteratively computes adaptation from source to target via backward hopping and from target to source via forward passing. In addition, we align low-level features for adapting image-level color/texture via adversarial training. However, the detector that performs well on both domains is not ideal for the target domain. As such, in each cycle, domain diversity is enforced by two regularizations: 1) maximum entropy regularization on the source domain to penalize confident source-specific learning and 2) minimum entropy regularization on target domain to intrigue target-specific learning. Theoretical analysis of the training process is provided, and extensive experiments on challenging cross-domain object detection datasets have shown our approach's superiority over the state-of-the-art.

1 Introduction

Object detection is a fundamental problem in computer vision [1–5], which can be applied to many scenarios such as face and pedestrian detection [6] and self-driving cars [7]. However, due to the variations in shape and appearance, lighting conditions and backgrounds, a model trained on the source data might not perform well on the target—a problem known as *domain discrepancy*. A common

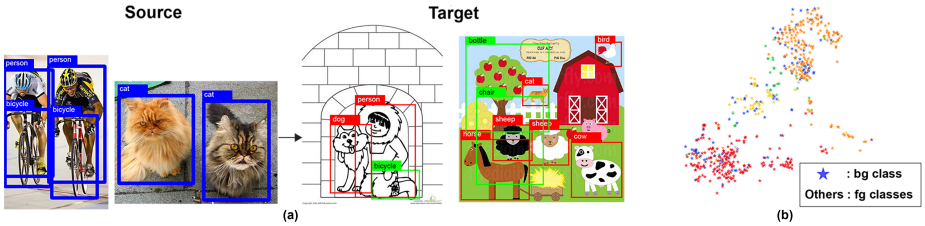

Fig. 1. (a) Due to domain discrepancy, the detector trained on the source domain does not perform well on the target. Green boxes indicate false positives and red indicate missing objects. (b) Feature visualization of the detection results on target images generated by source-only model. It is difficult to align feature at instance-level without category information due to the existence of false detections on the background.

approach to maximizing the performance on the target domain is via fine-tuning a pre-trained model with a large amount of target data. However, annotating bounding boxes for target objects is time-consuming and expensive. Hence, unsupervised domain adaptation methods for object detection are highly desirable.

Unsupervised domain adaptation for image classification has been extensively studied [8–13]. Most methods are developed to learn domain-invariant features by simultaneously minimizing the source error and the domain discrepancy through feature distribution alignment. Standard optimization criteria include maximum mean discrepancy [8, 9] and distribution moment matching [14, 15]. Recent adversarial training based methods have shown their effectiveness in learning domain-invariance by matching the marginal distributions of both source and target features [10, 16, 11]. However, this does not guarantee the alignment of class conditional distributions [17–20]. For example, aligning the target cat class to the source dog class can easily meet the objective of reducing the cost of source/target domain distinction, but the semantic categories are wrong. The limitation of adversarial learning is more evident when the domain discrepancy between two domains is larger, such as in object detection.

In object detection, performing domain alignment is more challenging compared to alignment in the image classification task in the following two aspects: (1) the input image may contain multiple objects, while there is only one centered object in the classification task; (2) the images in object detection are dominated by background and non-objects. Therefore, performing global adversarial learning (*i.e.,* marginal feature distributions) at the image-level is not sufficient for such challenging tasks due to the limitations discussed above. Chen *et al.* [21] made the first attempt to apply adversarial domain alignment to object detection, where the marginal feature distributions were aligned at both image-level and instance-level. However, due to the domain shift, the detector may not be accurate and many non-object proposals from the backgrounds are used for domain alignment (Fig. 1). This amplifies the limitation of adversarial domain training and hence limited gains can be achieved.

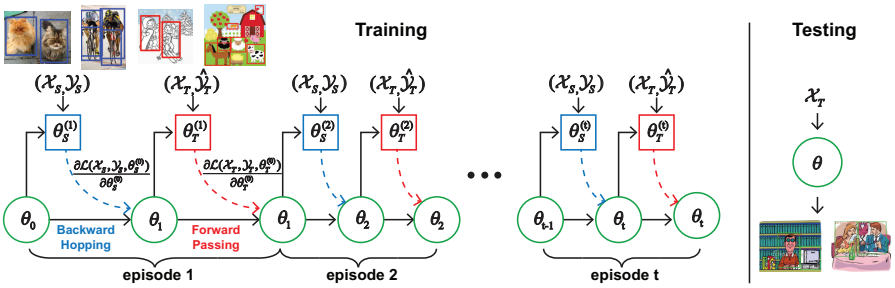

Fig. 2. The diagram of the proposed forward-backward cyclic adaptation for unsupervised domain adaptive object detection. In each episode, the training proceeds to achieve two goals: 1) gradient alignment across the source $\mathcal{X}_s$ and target $\mathcal{X}_t$ to achieve domain invariant detectors; and 2) encouraging domain-diversity to boost the target detector performance.

To tackle the limitation, efforts have been made to improve the image-level adaptation [22] and instance-level adaptation [23, 24] respectively. Saito *et al.* [22] proposed to weakly align the image-level features from the high-level layer, where the globally similar images have higher priorities to be aligned. In instance-level adaptation, Zhu *et al.* [23] proposed to filter the non-objects via grouping and then select source-like target instances according to the scores of the domain classifier. Zhuang *et al.* [24] proposed category-aware domain discriminators for instance-level alignment, where each category has its own domain discriminator.

We argue that explicit feature distribution alignment is not a necessary condition to learn domain-invariance. Instead, we remark that *domain-invariance* of category-level semantics can be learned by gradient alignment, where the inner product between the gradients of category-level classification loss from different domains is maximized. Intuitively, if the inner product is positive, taking a gradient step at the examples from one domain can decrease the loss at the examples from another domain. In other words, the learning of one domain can improve the learning of another domain and therefore lead to domain-invariance. More importantly, the gradients of category-level classification loss can encode class conditional information. Therefore, gradient alignment shows its advantages on the challenging instance-level adaptation for object detection.

In this work, we propose a Forward-Backward Cyclic Adaptation (FBC) approach to learn adaptive object detectors. In each cycle, the games of *Forward Passing*, an adaptation from source to target, and *Backward Hopping*, an adaptation from target to source, are played sequentially. Each adaptation is a domain transfer, where the training is first initialized with the model trained on the previous domain and then finetuned with the images in the current domain. We provide a theoretical analysis to show that by computing the forward and backward adaptation sequentially via Stochastic Gradient Descent (SGD), gradient alignment can be achieved. Our proposed approach is also related to the cycle consistency utilized in both machine translation [25] and image-to-image translation [26, 27] with a similar intuition that the mappings of an example transferred

from source to target and then back to the source domain should have the same results. In addition to instance-level adaptation via gradient alignment, we leverage adversarial domain training for image-level adaptation. Low-level features are aligned to learn the domain-invariance of holistic color and textures.

However, a detector with good generalization on both domains may not be the optimal solution for the target domain. To address this, we introduce *domain-diversity* into the training objective to avoid overfitting on the source domain and encourage target-specific learning on the target domain. We adopt two regularizers: (1) a maximum entropy regularizer on source domain and (2) a minimum entropy regularizer on the target domain.

We conduct experiments on four domain-shift scenarios and experimental results show the effectiveness of our proposed approach. **Contributions:** (1) We propose a forward-backward cyclic adaptation approach to learn unsupervised domain adaptive object detectors through image-level adaptation via adversarial domain alignment and instance-level adaptation via gradient alignment; (2) The proposed gradient alignment effectively aligns category-level semantics at the instance-level; (3) To achieve good performance on the target domain, we explicitly enforce domain-diversity via entropy regularization to further approximate the domain-invariant detectors closer to the optimal solution to target space; (4) The proposed method is simple yet effective and can be applied to various architectures.

2 Related Work

Object Detection. Deep object detection methods [28, 1, 3, 2, 5, 29] can be roughly grouped into two-stage detectors, *e.g.,* Faster R-CNN [1] and single-stage detectors, *e.g.,* SSD [2] and YOLO [3]. Faster R-CNN consists of two networks: a region proposal network and an R-CNN that classifies the proposals. Other methods like FPN [5] and RetinaNet [29] proposed to leverage a combination of features from different levels to improve the feature representations.

Unsupervised Domain Adaptation for Image Classification. A vast number of deep learning based unsupervised domain adaptation methods are presented for image classification. Many adaptation methods [8, 9, 14, 30, 10, 16, 11] are proposed to reduce the domain divergence based on the following theory:

Theorem 1 (Ben-David *et al.* [31]). *Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis in the hypothesis space $\mathcal{H}$. The expected error on target domain $\epsilon_T(h)$ is bounded by*

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda, \forall h \in \mathcal{H} \ , \tag{1}$$

where $\epsilon_S(h)$ is the expected error on the source domain,
$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h,h' \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S} [h(x) \neq h'(x)] - \Pr_{x \sim \mathcal{D}_T} [h(x) \neq h'(x)] \right|$ *measures domain divergence, and λ is the expected error of ideal joint hypothesis, $\lambda = \min_{h \in \mathcal{H}} [\epsilon_S(h) + \epsilon_T(h)]$.*

To minimize the divergence, various methods have been proposed to align the distributions of features from source and target domains, *e.g.,* maximum mean discrepancy [8, 9], correlation alignment [14], joint distribution discrepancy loss [30]

and adversarial training [10, 16, 11]. Adversarial training based methods [10, 16, 11] align the marginal distributions of source and target features, where the feature generator is trained to confuse the domain classifier. Although these methods have demonstrated impressive results, recent works [17, 22, 18, 32, 33] have shown that aligning marginal distributions without considering class conditional distributions does not guarantee small $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$. To address this, Luo *et al.* [19] proposed a semantic-aware discriminator and Xie *et al.* [17] proposed to align the semantic prototypes for each class. Some works [17, 32, 33] proposed to minimize the joint hypothesis error λ with pseudo labels in addition to the marginal distribution alignment. Some other methods proposed to use the predictions of a classifier as pseudo labels for unlabeled target samples [34, 12, 35]. Lee *et al.* [36] argued that training with pseudo labels is equivalent to entropy regularization, which favors a low-density separation between classes.

Unsupervised Domain Adaptation for Object Detection. Domain adaptive object detection has received much attention in the past two years [21, 23, 22, 37–40, 24]. The DA-Faster [21] was proposed to align domains at both image-level and instance-level by adding two domain classifiers to the Faster R-CNN. However, due to the limitation of domain adversarial training and inaccurate instance predictions, the improvement is limited. To improve the efficiency of image-level adaptation, multi-feature alignment [22, 37, 39, 24] has been proposed. In strong-weak domain alignment (SWDA) [22], Saito *et al.* proposed to strongly align low-level image features and weakly align high-level image features. Through weak alignment, the target images that are globally similar to source images have higher priorities to be aligned. Focal loss [29] is used in the domain classifier to achieve it. To address the inaccurate instance problem in instance-level adaptation, Zhu *et al.* [23] proposed to first filter non-object instances via grouping and then emphasize the target instances that are similar to the source for adversarial domain alignment. However, the category-level semantics are not studied in the traditional adversarial alignment. Zhuang *et al.* [24] proposed image-instance full alignment (iFAN) for category-aware instance-level adaptation, where each category owns a domain discriminator. Unlike using adversarial training, our proposed method aligns category-level semantics via gradient alignment.

Gradient-based Meta Learning and Continual Learning. Our method is also related to recent gradient-based meta-learning methods: MAML [41] and Reptile [42], which are designed to learn a good initialization for few-shot learning and have demonstrated good within-task generalization. Reptile [42] suggested that SGD automatically maximizes the inner products between the gradients computed on different minibatches of the same task, and results in within-task generalization. Riemer *et al.* [43] integrated the Reptile with an experience replay module for the task of continual learning, where the transfer between examples is maximized via gradient alignment. Inspired by these methods, we leverage the generalization ability of Reptile [42] to improve the generalization across domains for unsupervised domain adaptation via gradient alignment.

Entropy Regularization. The maximum entropy principle proposed by Jaynes [44] has been applied to reinforcement learning [45, 46] to prevent early convergence

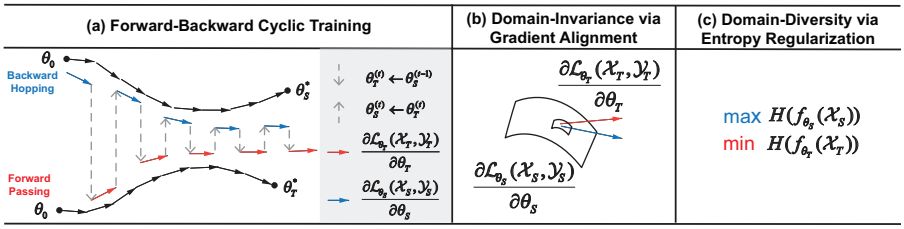

Fig. 3. (a) Illustration of the model updates in our proposed forward-backward cyclic adaptation method. The θ_0 is the initial model and the θ_S^* and θ_T^* are the optimal solutions for source and target domain, respectively. (b) We propose that domain-invariance occurs when the gradients of source and target samples are pointing in similar directions. (c) The domain diversity is implemented by maximum entropy regularization on the source domain and minimum entropy regularization on the target domain.

and supervised learning to improve generalization [47–50]. On the contrary, the entropy minimization has been used for unsupervised clustering [51], semi-supervised learning [52] and unsupervised domain adaptation [9, 53] to encourages low density separation between clusters or classes.

3 Forward-Backward Domain Adaptation for Object Detection

3.1 Overview

In unsupervised domain adaptation, N_S labeled images $\{\mathcal{X}_S, \mathcal{Y}_S\} = \{x_S^i, y_S^i\}_{i=1}^{N_S}$ from the source domain with a distribution $\mathcal{D}_S$ are given. We have N_T unlabeled images $\mathcal{X}_T = \{x_T^j\}_{j=1}^{N_T}$ from the target domain with a different distribution $\mathcal{D}_T$, but the ground truth labels $\mathcal{Y}_T = \{y_T^j\}_{j=1}^{N_T}$ are not accessible during training. Note that in object detection, each label in $\mathcal{Y}_S$ or $\mathcal{Y}_T$ is composed of a set of bounding boxes with their corresponding class labels. Our goal is to learn a neural network (parameterized by θ) $f_\theta : \mathcal{X}_T \rightarrow \mathcal{Y}_T$ that can make accurate predictions on the target samples without the need for labeled training data.

In this work, we argue that aligning the feature distributions is not a necessary condition to reduce the $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ in Theorem 1. Unlike the above-mentioned distribution alignment based methods, we cast the domain adaptation into an optimization problem to learn the domain-invariance. We propose to utilize gradient alignment for category-aware instance-level adaptation. For image-level adaptation, local feature alignment via adversarial training is performed. As the ultimate goal of domain adaptation is to achieve good performance on the target domain, we further introduce *domain-diversity* into training to boost the detection performance in the target space.

3.2 Gradient Alignment via Forward-Backward Cyclic Training

Recent gradient-based meta-learning methods [41, 54, 42], designed for few-shot learning, have demonstrated their success in approximating learning algorithms

and shown their ability to generalize well to new data from unseen distributions. Inspired by these methods, we propose to learn the *domain-invariance* via gradient alignment to achieve generalization across domains.

Gradient Alignment for Domain-invariance Suppose that we have neural networks that learn the predictions for source and target samples as $f_{\theta_S} : \mathcal{X}_S \to \mathcal{Y}_S$ and $f_{\theta_T} : \mathcal{X}_T \to \mathcal{Y}_T$. The network parameters θ_S and θ_T are updated via minimizing the empirical risks, $\mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S) = \frac{1}{N_S}\sum_{i=1}^{N_S} \ell(f_{\theta_S}(x_S^i), y_S^i)$ and $\mathcal{L}_{\theta_T}(\mathcal{X}_T, \mathcal{Y}_T) = \frac{1}{N_T}\sum_{j=1}^{N_T} \ell(f_{\theta_T}(x_T^j), y_T^j)$, where $\ell(\cdot)$ is the cross-entropy loss. Inspired by methods for continual learning [43, 55], when the parameters θ_S and θ_T are shared and the gradient updates are in small steps, we could assume the function $\mathcal{L}_\theta$ is linear. If the following condition is satisfied, the gradient updates at source samples could decrease the loss at target samples and vice verse:

$$\frac{\partial \mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S)}{\partial \theta_S} \cdot \frac{\partial \mathcal{L}_{\theta_T}(\mathcal{X}_T, \mathcal{Y}_T)}{\partial \theta_T} > 0 \ , \tag{2}$$

where the $\cdot$ is the inner-product operator. This indicates that the learning of one domain could improve the learning of another domain. Therefore, we propose that domain-invariance could be learned by maximizing the inner products of gradients from different domains. Moreover, this gradient alignment can encode category-level semantics as the gradients are generated from the classification losses $\mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S)$ and $\mathcal{L}_{\theta_T}(\mathcal{X}_T, \mathcal{Y}_T)$. It is different from the feature alignment by a domain classifier in adversarial training based methods [10, 16, 11, 21, 22], where class information is not explicitly considered. Thus, we use gradient alignment for instance-level adaptation.

Recall Theorem 1, once $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is minimized, the generalization error on target domain $\epsilon_T(h)$ is bounded by the shared error of ideal joint hypothesis, $\lambda = \min_{h\in\mathcal{H}}[\epsilon_S(h) + \epsilon_T(h)]$. As suggested in [31], it is important to have a classifier performing well on both domains. Therefore, similar to the previous works [36, 17, 33], we resort to using pseudo labels $\hat{\mathcal{Y}}_T = \{\hat{y}_T^j\}_{j=1}^{N_T}$ to optimize the upper bound for the λ. These pseudo labels are the detections on the target images produced by the source detector f_{θ_S} and are updated with the updates of f_{θ_S}. Our objective function of gradient alignment is to minimize the $\mathcal{L}_g$:

$$\mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S) + \mathcal{L}_{\theta_T}(\mathcal{X}_T, \hat{\mathcal{Y}}_T) - \alpha \frac{\partial \mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S)}{\partial \theta_S} \cdot \frac{\partial \mathcal{L}_{\theta_T}(\mathcal{X}_T, \hat{\mathcal{Y}}_T)}{\partial \theta_T} \ . \tag{3}$$

Forward-Backward Cyclic Training To achieve the above objective, we propose an algorithm that sequentially plays the game of *Backward Hopping* on the source domain and *Foward Passing* on the target domain, and a shared network parameterized by θ is updated iteratively. We initialize the shared network θ with ImageNet [56] pre-trained model. Let us denote a cycle of performing forward passing and backward hopping as an episode. In the backward hopping phase of episode t, the network parameterized by $\theta_S^{(t)}$ is first initialized with the model $\theta_T^{(t-1)}$ from the previous episode $t-1$. And the model $\theta_S^{(t)}$ is then optimized with one image per time via stochastic gradient descent (SGD) on N_S labeled source

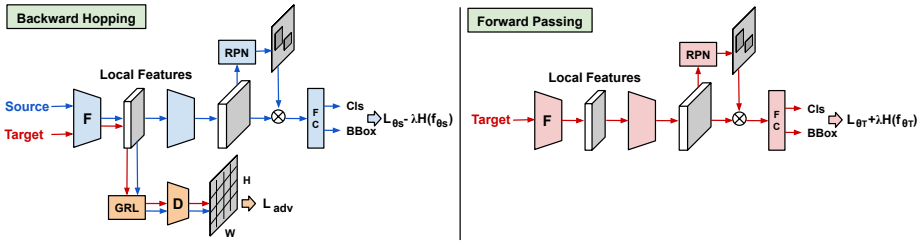

Fig. 4. Network Architecture.

images $\{\mathcal{X}_S, \mathcal{Y}_S\}$. In forward passing, the model $\theta_T^{(t)}$ is initialized with $\theta_S^{(t)}$ and trained with pseudo labeled target samples $\{\mathcal{X}_T, \hat{\mathcal{Y}}_T\}$. The training procedure is shown in Fig. 2.

Theoretical Analysis. We provide a theoretical analysis to show how our proposed forward and backward training strategy can achieve the objective of gradient alignment in Eq. 3. For simplicity, we only analyze the gradient computations in one episode and denote the gradient obtained in one episode as g_e. We then have $g_e = g_S + g_T$, where g_S is obtained in backward hopping $g_S = \frac{\partial \mathcal{L}_{\theta_S}(\mathcal{X}_S, \mathcal{Y}_S)}{\partial \theta_S}$ and g_T is the gradient obtained in forward passing $g_T = \frac{\partial \mathcal{L}_{\theta_T}(\mathcal{X}_T, \hat{\mathcal{Y}}_T)}{\partial \theta_T}$.

According to Taylor's theorem, the gradient of forward passing can be expanded as $g_T = \bar{g}_T + \bar{H}_T(\theta_T - \theta_0) + O(\|\theta_T - \theta_0\|^2)$, where $\bar{g}_T$ and $\bar{H}_T$ are the gradient and Hessian matrix at initial point θ_0. Then the overall gradient g_e can be rewritten as:

$$g_e = g_S + g_T = \bar{g}_S + \bar{g}_T + \bar{H}_T(\theta_T - \theta_0) + O(\|\theta_T - \theta_0\|^2) \ . \tag{4}$$

Let us denote the initial parameters in one episode as θ_0. In our proposed forward and backward training strategy, the model parameters of backward hopping are first initialized with $\theta_S = \theta_0$ and are updated by $\theta_0 - \alpha g_S$. In forward passing, the model is initialized with the updated θ_S and thus $\theta_T = \theta_0 - \alpha g_S$. Substitute this to Eq. 4 and we have

$$g_e = \bar{g}_S + \bar{g}_T - \alpha \bar{H}_T \bar{g}_S + O(\|\theta_T - \theta_0\|^2) \ . \tag{5}$$

It is noted in Reptile [42] that $\mathbb{E}[\bar{H}_S \bar{g}_T] = \mathbb{E}[\bar{H}_T \bar{g}_S] = \frac{1}{2}[\frac{\partial}{\partial \theta_0}(\bar{g}_S \cdot \bar{g}_T)]$. Therefore, this training is approximating our objective function in Eq. 3. More details are shown in the supplementary materials.

3.3 Local Feature Alignment via Adversarial Training

Domain adversarial training has demonstrated its effectiveness in reducing domain discrepancy of low-level features, *e.g.,* local texture and color, regardless of class conditional information [21, 22]. Therefore, we align the low-level features at the image-level in combination with gradient alignment on the source domain. We utilize the gradient reversal layer (GRL) proposed by Ganin and Lempitsky [10] for adversarial domain training, where the gradients of the domain classifier are reversed for domain confusion. Following SWDA [22], we extract

local features F from a low-level layer as the input of the domain classifier D and the least-squares loss [57, 26] is used to optimize the domain classifier. The loss of adversarial training is as follows:

$$\mathcal{L}_{adv} = \frac{1}{2}\frac{1}{N_S WH}\sum_{i,w,h} D(F(x_S^i))_{wh}^2 + \frac{1}{2}\frac{1}{N_S WH}\sum_{j,w,h}(1 - D(F(x_T^j))_{wh})^2 \ , \quad (6)$$

where H and W are the height and width of the output feature map of the domain classifier.

3.4 Domain Diversity via Entropy Regularization

The ultimate goal of domain adaptation is to achieve good performance on the target domain. However, a model that only learns the domain-invariance is not an optimal solution for the target domain, as

$$\epsilon_T(h) \leq \epsilon_T(h^a) + \epsilon_T(h, h^a), \quad (7)$$

where $h^a = \arg\min_{h\in\mathcal{H}}[\epsilon_S(h) + \epsilon_T(h)]$. Moreover, in the absence of ground truth labels for target samples, the learning of domain-invariance largely relies on the source samples, which might lead to overfitting on the source domain and limiting its ability to generalize well on target domain. Therefore, it is crucial to introduce the domain-diversity into the training to encourage more emphasis on target-specific information.

We define the domain diversity as a combination of two regularizations: (1) maximum entropy regularization on the source domain to avoid overfitting and (2) minimum entropy regularization on unlabeled target domain to leverage target-specific information. Low entropy corresponds to high confidence. To avoid the overfitting when training with source data, we utilize the maximum entropy regularizer [47] to penalize the confident predictions with low entropy:

$$\max_{\theta_S} \mathrm{H}(f_{\theta_S}(\mathcal{X}_S)) = -\sum_{i=1}^{N_S} f_{\theta_S}(x_S^i)\log(f_{\theta_S}(x_S^i)) \ . \quad (8)$$

On the contrary, to leverage unlabeled target domain data, we exploit the minimum entropy regularizer. The entropy minimization has been used for unsupervised clustering [51], semi-supervised learning [52] and unsupervised domain adaptation [9, 53] to encourages low density separation between clusters or classes. Here, we minimize the entropy of class conditional distribution:

$$\min_{\theta_T} \mathrm{H}(f_{\theta_T}(\mathcal{X}_T)) = -\sum_{j=1}^{N_T} f_{\theta_T}(x_T^j)\log(f_{\theta_T}(x_T^j)) \ . \quad (9)$$

We define the objective of domain diversity is to minimize the following function:

$$\mathcal{L}_{div}(\mathcal{X}_S, \mathcal{X}_T) = -\mathrm{H}(f_{\theta_S}(\mathcal{X}_S)) + \mathrm{H}(f_{\theta_T}(\mathcal{X}_T)) \ . \quad (10)$$

3.5 Overall Objective

To learn domain-invariance for adapting object detectors, we perform gradient alignment for high-level semantics and domain adversarial training on local features for low-level information. The loss function of domain-invariance is:

$$\mathcal{L}_{inv}(\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T) = \mathcal{L}_g(\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T) + \lambda\mathcal{L}_{adv}(\mathcal{X}_S, \mathcal{X}_T) \ , \quad (11)$$

Algorithm 1 Forward-Backward Cyclic Domain Adaptation for Object Detection

Input: Source samples $\{x_S^i, y_S^i\}_{i=1}^{N_S}$, target samples $\{x_T^j\}_{j=1}^{N_T}$, ImageNet pre-trained model θ_0, hyperparameters α, β, γ, λ, number of iterations N_{itr}

Output: A shared model θ

Initialize θ with θ_0
for t in N_{itr} **do**
  //Backward Hopping:
  $\theta_S^{(t)} \leftarrow \theta$
  for i, j in N_S, N_T **do**
    $\theta_S^{(t)} \leftarrow \theta_S^{(t)} - \alpha \nabla_{\theta_S^{(t)}} (\mathcal{L}_{\theta_S^{(t)}}(x_S^i, y_S^i) + \lambda \mathcal{L}_{adv}(x_S^i, x_T^j) - \gamma \mathrm{H}(f_{\theta_S^{(t)}}(x_S^i)))$
  end for
  $\theta \leftarrow \theta - \beta \theta_S^{(t)}$
  Generate pseudo labels $\hat{y}_T = f_{\theta_S^{(t)}}(x_T^j), j = 1, ..., N_T$
  //Forward Passing:
  $\theta_T^{(t)} \leftarrow \theta$
  for j in N_T **do**
    $\theta_T^{(t)} \leftarrow \theta_T^{(t)} - \alpha \nabla_{\theta_T^{(t)}} (\mathcal{L}_{\theta_T^{(t)}}(x_T^j, \hat{y}_T^j) + \gamma \mathrm{H}(f_{\theta_T^{(t)}}(x_T^j)))$
  end for
  $\theta \leftarrow \theta - \beta \theta_T^{(t)}$
end for

where λ balances the trade-off between gradient alignment loss and adversarial training loss.

Maximizing the domain-diversity contradicts the intention of learning domain-invariance. However, without access to the ground truth labels of target samples, the accuracy of the target samples relies on the domain-invariance information learned from the source domain. Consequently, it is important to accomplish the trade-off between learning domain-invariance and domain-diversity. We use a hyperparameter γ to balance the trade-off. Our overall objective function is

$$\min_{\theta} \mathcal{L}_{inv}(\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T) + \gamma \mathcal{L}_{div}(\mathcal{X}_S, \mathcal{X}_T) \ . \tag{12}$$

The full algorithm is outlined in Algorithm 1.

4 Experiments

In this section, we evaluate the proposed forward-backward cycling adaptation approach (FBC) on four cross-domain detection datasets.

4.1 Implementation Details

Following DA-Faster [21] and SWDA [22], we use the Faster-RCNN [1] as our detection framework. All training and test images are resized with the shorter side of 600 pixels and the training batch size is 1. Our method is implemented using Pytorch. The source only model is fine-tuned on the pre-trained ImageNet [56] model with labeled source samples without adaptation For the evaluation, we measure the mean average precision (mAP) with a threshold of 0.5 across all classes. More details are shown in supplementary materials.

Table 1. Results (%) on the adaptation from PASCAL [58] to Clipart Dataset [59]. The DA-Faster†is the reported result in SWDA [22].

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	prsn	plnt	sheep	sofa	train	tv	mAP
Source Only	24.2	47.1	24.9	17.7	26.6	47.3	30.4	11.9	36.8	26.4	10.1	11.8	25.9	74.6	42.1	24.0	3.8	27.2	37.9	29.9	29.5
DA-Faster†[21]	15.0	34.6	12.4	11.9	19.8	21.1	23.2	3.1	22.1	26.3	10.6	10.0	19.6	39.4	34.6	29.3	1.0	17.1	19.7	24.8	19.8
SWDA [22]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
FBC (ours)	43.9	64.4	28.9	26.3	39.4	58.9	36.7	14.8	46.2	39.2	11.0	11.0	31.1	77.1	48.1	36.1	17.8	35.2	52.6	50.5	**38.5**

4.2 Adaptation between Dissimilar Domains

We evaluate the adaptation performance on two pairs of dissimilar domains: PASCAL [58] to Clipart [59], and PASCAL [58] to Watercolor [59]. For the two domain shifts, we use the same source-only model trained on PASCAL. Following SWDA [22], we use ResNet101 [60] as the backbone network for Faster R-CNN detector and the settings of training and test sets are the same.

Datasets. PASCAL VOC dataset [58] is used as the source domain in these two domain shift scenarios. This dataset consists of real images with 20 object classes. The training set contains around 15K images. The two dissimilar target domains are Clipart dataset [59] with comic images and Watercolor dataset [59] with artistic images. Clipart dataset has the same 20 object classes as the PASCAL, while Watercolor only has six. Clipart dataset contains 1K comic images, which are used for both training (without labels) and testing. There are 2K images in the Watercolor dataset: 1K for training (without labels) and 1K for testing.

Results on the Clipart Dataset [59]. In the original paper of DA-Faster [21], they do not evaluate the Clipart and Watercolor datasets. Thus, we follow with the results of DA-Faster [21] reported in SWDA [22]. As shown in Table 1, in comparison to the source only model, DA-Faster [21] degrades the detection performance significantly, with a drop of 8 percentage points in mAP. DA-Faster [21] adopts two domain classifiers on both image-level and instance-level features. However, the source/target domain confusion without considering the semantic information will lead to the wrong alignment of semantic classes across domains. The problem is more challenging when domain shift in object detection is large, *i.e.,* PASCAL [58] to Clipart [59]. In Clipart, the comic images contain objects that are far different from those in PASCAL w.r.t. the shapes and appearance, such as sketches. To address this, the SWDA [22] conducts a weak alignment on the image-level features by training the domain classifier with a focal loss. With the additional help of a domain classifier on lower level features and context regularization, the SWDA [22] can boost the mAP of detection from 27.8% to 38.1% with an increase of 10.3 points. Our proposed FBC can achieve the highest mAP of 38.5%. In the ablation studies (Table 2), we can see that using gradient alignment only could also obtain good performance in this challenging adaptation scenario.

Results on the Watercolor Dataset [59]. The adaptation results are summarized in Table 3. In Watercolor, most of the images contain only one or two objects with less variations of shape and appearance than those in the Clipart. As reported in SWDA [22], the source only model can achieve quite good results with an mAP of 44.6% and DA-Faster [21] can improve it slightly by only 1.4 points. SWDA [22] performs much better than DA-Faster [21] and obtain a high

Table 2. Ablation studies of the proposed method on the adaptation from PASCAL [58] to Clipart Dataset [59]. G: gradient alignment, L: local feature alignment and D: domain diversity.

G	L	D	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	prsn	plnt	sheep	sofa	train	tv	mAP
✓			28.8	64	21.1	19.1	39.7	60.7	29.5	14.2	46.4	29.3	21.8	8.9	28.8	72.7	51.3	32.9	12.8	28.1	52.7	49.5	35.6
✓		✓	32.1	57.6	24.4	23.7	34.1	59.3	32.2	9.1	40.3	41.3	27.8	11.9	30.2	72.9	48.8	38.3	6.1	33.1	46.5	48	35.9
	✓		31.8	53.0	21.3	25.0	36.1	55.9	30.4	11.6	39.3	21.0	9.4	14.5	32.4	79.0	44.9	37.8	6.2	35.6	43.0	53.5	34.1
✓	✓	✓	43.9	64.4	28.9	26.3	39.4	58.9	36.7	14.8	46.2	39.2	11.0	11.0	31.1	77.1	48.1	36.1	17.8	35.2	52.6	50.5	**38.5**

Table 3. Results (%) on the adaptation from PASCAL [58] to Watercolor [59]. The DA-Faster†is the reproduced result in SWDA [21]. G: gradient alignment, L: local feature alignment and D: domain diversity.

Method	G	L	D	bike	bird	car	cat	dog	prsn	mAP
Source Only (ours)				66.7	43.5	41.0	26.0	22.9	58.9	43.2
DA-Faster† [21]				75.2	40.6	48.0	31.5	20.6	60.0	46.0
SWDA [22]				82.3	55.9	46.5	32.7	35.5	66.7	53.3
FBC (ours)	✓			90.0	46.5	51.3	33.2	29.5	65.9	52.9
	✓		✓	88.7	48.2	46.6	38.7	35.6	64.1	53.6
		✓		89.0	47.2	46.1	39.9	27.7	65.0	52.5
	✓	✓	✓	90.1	49.7	44.1	41.1	34.6	70.3	**55.0**

mAP of 53.3%. The gain from adaptation is 8.7 points. The mAP of our proposed FBC is 55.0%, which is 1.5% higher than that of SWDA. Even without the local feature alignment via adversarial training, our proposed forward-backward cyclic adaptation method (53.6%) can achieve state-of-the-art performance.

Feature Visualization. To visualize the adaptability of our method, we use the Grad-cam [61] to show the evidence (heatmap) for the last fully connected layer in the object detectors. The high value in the heatmap indicates the evidence why the classifiers make the classification. Fig. 5 shows the differences of classification evidence before and after adaptation. As we can see, the adapted detector is able to classify the objects (*e.g.,* persons) based on more semantics (*e.g.,* faces, necks, joints). It demonstrates that the adapted detector has addressed the discrepancy on the appearance of real and cartoon objects.

4.3 Adaptation from Synthetic to Real Images

As the adaptation from the synthetic images to the real images can potentially reduce the efforts of collecting the real data and labels, we evaluate the adaptation performance in the scenario of Sim10k [62] to Cityscapes [63].

Datasets. The source domain, Sim10k [62], contains synthetic images that are rendered by the computer game Grand Theft Auto (GTA). It provides 58,701 bounding box annotations for cars in 10K images. The target domain, Cityscapes [63], consists of real images captured by car-mounted video cameras for driving scenarios. It comprises 2,975 images for training and 500 images for validation. We use its training set for adaptation without labels and validation set for evaluation. The adaptation is only evaluated on class *car* as Sim10k only provides annotations for car.

Results. Results are shown in Table 4. The reported mAP gain of DA-Faster [21] in its original report (7.8 points) is significantly different from its reproduced gain (-0.4 points) in SWDA [22]. It implies that a lot of efforts are needed to reproduce the reported results of DA-Faster [21]. Our proposed FBC has a competitive

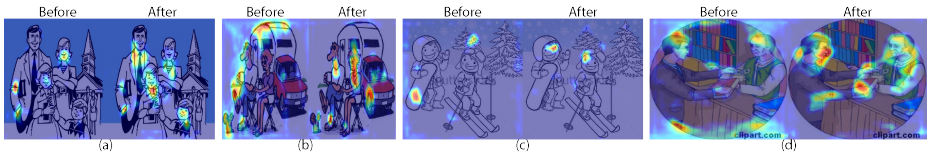

Fig. 5. Feature visualization for showing the evidence for classifiers before and after domain adaptation using Grad-cam [61].

Table 4. Results (%) on the adaptation from Sim10k [62] to Cityscapes [63]. The DA-Faster†is the reproduced result in SWDA [21]. G: gradient alignment, L: local feature alignment and D: domain diversity.

Method	G	L	D	AP on Car
Source Only (ours)				31.2
DA-Faster [21]				39.0
DA-Faster† [21]				34.2
MAF [37]				41.1
SWDA [22]				42.3
Zhu *et al.* [23]				43.0
iFAN [24]				**46.2**
FBC (ours)	✓			38.2
	✓		✓	39.2
		✓		41.4
	✓	✓	✓	42.7

result of mAP, 42.7%, which is 0.4% higher than that of SWDA and on par with that of Zhu *et al.* (43 %). iFAN *et al.* [24] achieve the best performance with an mAP of 46.2%. We note that for image-level adaptation, iFAN adopts four domain classifiers for aligning multi-level features, whereas we only align the features from a single layer. Despite this, our proposed method could obtain better results than iFAN in the adaptation from Cityscapes to FoggyCityscapes.

4.4 Adaptation between Similar Domains

Datasets. The target dataset, FoggyCityscapes [64], is a synthetic foggy dataset where images are rendered from the Cityscapes [63]. The annotations and data splits are the same as the Cityscapes. The adaptation performance is evaluated on the validation set of FoggyCityscapes.

Results. It can be seen in Table 5 that our proposed FBC method outperforms the baseline methods, which boosts the mAP to 36.7%. It is noteworthy that MAF (34.0%), iFAN (35.5 %) and Xie *et al.* (36.0 %) utilize multiple domain classifiers for multi-layer image-level feature alignment, whereas we only use single-layer features. If without the local feature alignment, our proposed method can only obtain limited gain. It is because, in this scenario, the main difference between these two domains is the local texture. But with the combination of gradient alignment and domain diversity, our full model could achieve state-of-the-art performance.

t-SNE Visualization. We visualize the differences of features before and after adaptation via t-SNE visualization [65] in Fig. 6. The features are output from the ROI pooling layer and 100 images are randomly selected. After adaptation,

Table 5. Results (%) on the adaptation from Cityscapes [63] to FoggyCityscapes Dataset [64]. G: gradient alignment, L: local feature alignment and D: domain diversity.

Method	G	L	D	prsn	rider	car	truck	bus	train	motor	bcycle	mAP
Source Only (ours)				22.4	34.2	27.2	12.1	28.4	9.5	20.0	27.1	22.9
DA-Faster [21]				25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
Zhu *et al.* [23]				33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MAF [37]				28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SWDA [22]				29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
Diversify&Match [38]				30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
iFAN [24]				32.6	48.5	22.8	40.0	33.0	45.5	31.7	27.9	35.3
Xie *et al.* [39]				33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
FBC (ours)	✓			25.8	35.6	35.5	18.4	29.6	10.0	24.5	30.3	26.2
	✓		✓	29.0	37.0	35.6	18.9	32.1	10.7	25.0	31.3	27.5
		✓		31.6	45.1	42.6	26.4	37.8	22.1	29.4	34.6	33.7
	✓	✓	✓	31.5	46.0	44.3	25.9	40.6	39.7	29.0	36.4	**36.7**

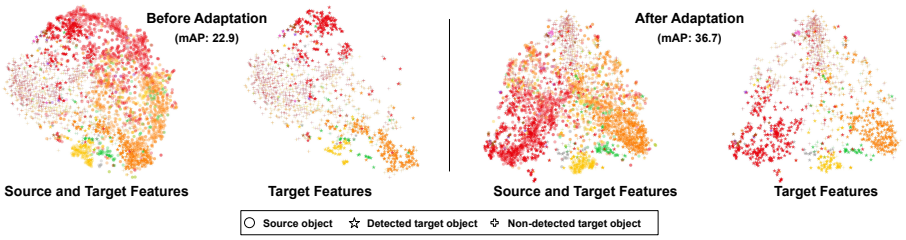

Fig. 6. t-SNE visualization of features before and after domain adaptation from Cityscape to FoggyCityScape. Different colors represent different classes. Target features are displayed alone on the right for better visualization.

the distributions of source and target features are well aligned with regard to the object classes. More importantly, as shown in Fig.6, different classes are better distinguished and more target objects are detected for each class after adaptation. This demonstrates the effectiveness of our proposed adaptation method for object detection.

5 Conclusions

We address unsupervised domain adaptation for object detection task where the target domain does not have labels. A forward-backward cyclic adaptation method is proposed. This method was based on the intuition that domain invariance of category-level semantics could be learned when the gradient directions of source and target were aligned. Theoretical analysis was presented to show that the proposed method achieved the gradient alignment goal. Local feature alignment via adversarial training was performed for learning domain-invariance of holistic color/textures. Furthermore, we proposed a domain diversity constraint to penalize confident source-specific learning and intrigue target-specific learning via entropy regularization.

Acknowledgements. This research was funded by the Australian Government through the Australian Research Council and Sullivan Nicolaides Pathology under Linkage Project LP160101797. Lin Wu was supported by NSFC U19A2073, the Fundamental Research Funds for the Central Universities under Grant No.JZ2020HGTB0050.

References

1. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. (2015)
2. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. (2016)
3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. (2016)
4. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
5. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. (2017)
6. Hu, P., Ramanan, D.: Finding tiny faces. In: CVPR. (2017)
7. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: CVPR. (2017)
8. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. ICML (2015)
9. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: NeurIPS. (2016)
10. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. (2015)
11. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. (2017)
12. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. In: ICML. (2017)
13. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. (2018)
14. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV. (2016)
15. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: ICCV. (2015)
16. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
17. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: ICML. (2018)
18. Shu, R., Bui, H.H., Narui, H., Ermon, S.: A dirt-t approach to unsupervised domain adaptation. In: ICLR. (2018)
19. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. arXiv preprint arXiv:1809.09478 (2018)
20. Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. In: NeurIPS. (2018)
21. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. (2018)
22. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: CVPR. (2019)
23. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: CVPR. (2019)

24. Zhuang, C., Han, X., Huang, W., Scott, M.R.: ifan: Image-instance full alignment networks for adaptive object detection. In: AAAI. (2020)
25. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.Y., Ma, W.Y.: Dual learning for machine translation. In: NeurIPS. (2016)
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. (2017)
27. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV. (2017)
28. Girshick, R.: Fast r-cnn. In: ICCV. (2015)
29. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: CVPR. (2017)
30. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML. (2017)
31. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79** (2010) 151–175
32. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. (2019)
33. Chen, C., Xie, W., Xu, T., Huang, W., Rong, Y., Ding, X., Huang, Y., Huang, J.: Progressive feature alignment for unsupervised domain adaptation. In: CVPR. (2019)
34. Sener, O., Song, H.O., Saxena, A., Savarese, S.: Learning transferrable representations for unsupervised domain adaptation. In: NeurIPS. (2016)
35. Chen, M., Weinberger, K.Q., Blitzer, J.: Co-training for domain adaptation. In: NeurIPS. (2011)
36. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. Volume 3. (2013) 2
37. He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: ICCV. (2019)
38. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: CVPR. (2019)
39. Xie, R., Yu, F., Wang, J., Wang, Y., Zhang, L.: Multi-level domain adaptive learning for cross-domain detection. In: ICCV Workshops. (2019)
40. Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: The IEEE Winter Conference on Applications of Computer Vision. (2020)
41. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. (2017)
42. Nichol, A., Schulman, J.: Reptile: a scalable metalearning algorithm. arXiv preprint arXiv:1803.02999 **2** (2018)
43. Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G.: Learning to learn without forgetting by maximizing transfer and minimizing interference. ICLR (2019)
44. Jaynes, E.T.: Information theory and statistical mechanics. Physical review (1957)
45. Williams, R.J., Peng, J.: Function optimization using connectionist reinforcement learning algorithms. Connection Science (1991)
46. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: ICML. (2016) 1928–1937

47. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)
48. Liu, H., Jin, S., Zhang, C.: Connectionist temporal classification with maximum entropy regularization. In: NeurIPS. (2018)
49. Dubey, A., Gupta, O., Raskar, R., Naik, N.: Maximum-entropy fine grained classification. In: NeurIPS. (2018)
50. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 568–583
51. Palubinskas, G., Descombes, X., Kruggel, F.: An unsupervised clustering method using the entropy minimization. In: ICPR. (1998)
52. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NeurIPS. (2005)
53. Luo, Z., Zou, Y., Hoffman, J., Fei-Fei, L.F.: Label efficient learning of transferable representations acrosss domains and tasks. In: NeurIPS. (2017)
54. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. ICLR (2016)
55. Lopez-Paz, D., et al.: Gradient episodic memory for continual learning. In: NeurIPS. (2017) 6467–6476
56. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
57. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. (2017)
58. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88** (2010) 303–338
59. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: CVPR. (2018)
60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
61. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. (2017)
62. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)
63. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. (2016)
64. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. IJCV (2018) 1–20
65. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008) 2579–2605