# Category Dictionary Guided Unsupervised Domain Adaptation for Object Detection

**Shuai Li,**[1,2] **Jianqiang Huang,**[2] **Xian-Sheng Hua,**[2] **Lei Zhang**[1,2]
[1]The Hong Kong Polytechnic University
[2]DAMO Academy, Alibaba Group
{csshuaili,cslzhang}@comp.polyu.edu.hk, {jianqiang.hjq, xiansheng.hxs}@alibaba-inc.com

## Abstract

Unsupervised domain adaption (UDA) is a promising solution to enhance the generalization ability of a model from a source domain to a target domain without manually annotating labels for the target data. Recent works in cross-domain object detection mostly resort to adversarial feature adaptation to match the marginal distributions of two domains. However, perfect feature alignment is hard to achieve and what's more is likely to cause negative transfer due to the high complexity of object detection. In this paper, we take a different approach to reduce the domain gap by a self-training paradigm, which regards the pseudo-labels as ground truth to fully exploit the unlabeled target data. In order to generate more informative pseudo labels, we further propose a category dictionary guided (CDG) UDA model for cross-domain object detection, which learns category-specific dictionaries from the source domain to represent the candidate boxes in target domain. The representation residual can be used for not only pseudo label assignment but also quality (e.g., IoU) estimation of the candidate box. Compared with decision boundary based classifiers such as softmax, the proposed CDG scheme can select more informative and reliable pseudo-boxes. Experimental results on benchmark datasets show that the proposed CDG significantly exceeds the state-of-the-arts in cross-domain object detection.

## Introduction

Object detection is a fundamental computer vision task, aiming to detect object labels and locations in an image. The detection performance has gone through a continuous growth due to the rapid development of deep learning techniques (Simonyan and Zisserman 2014; He et al. 2016) and large scale datasets (Lin et al. 2014; Deng et al. 2009) in recent years. It is often assumed that training data and test data follow the same distribution, which however may not be true in real world environments. Factors such as illumination, viewpoints, weather condition and cameras can cause domain shifts between the source data and target data. A state-of-the-art detector trained on the source data may degrade drastically when applied to the target data. One possible solution is to annotate new data and retrain the model. Unfortunately, annotating labels at the box level is expensive and time-consuming. Another promising solution is domain adaptation (DA), which aims to minimize the gap between the two domains and learn a shared discriminative model for both domains (Ganin and Lempitsky 2014; Long et al. 2016; Tzeng et al. 2017; Xie et al. 2018; Lee et al. 2019; Xu et al. 2019). In this paper, we focus on the challenging unsupervised domain adaptation (UDA) problem, where no labels are available for target domain.

Many previous UDA methods minimize the domain shift by performing feature alignment to learn domain-invariant features. Following this line of research, most cross-domain object detection methods attempt to address domain adaptation by minimizing the domain discrepancy at different levels, such as appearance level (Inoue et al. 2018) and feature level (Cai et al. 2019; Khodabandeh et al. 2019; Saito et al. 2019; Zhu et al. 2019). Compared with classification and segmentation, feature alignment for detection is much harder given the complex combinations of various objects and different scene layouts between the two domains. Brutally enforcing global feature alignment may cause negative transfer and hurt the discrimination ability of the final detector. In addition, learning domain-invariant features ignores the domain-specific knowledge of target data, which is of particular importance for detection as the target domain has its special characteristics that cannot be well encoded in the aligned feature space.

Another line of research focuses on learning domain specific knowledge by fully exploiting the target data. The domain specific knowledge has been explored in classification (Zou et al. 2019) and segmentation (Zou et al. 2018; Zhang et al. 2019) by utilizing a self-training paradigm on the pseudo-labeled target data annotated by the knowledge from source domain (e.g., the model trained on source data). Self-training is one of the powerful ways for domain adaptation since it can achieve class-wise adaptation. In this line of research, how to generate high quality pseudo labels is critical to the final performance as error-prone labels may mislead the detector training. It is required that the assigned pseudo boxes should be reliable and informative. The classifier used in the detector, e.g., softmax classifier, is not effective to assign labels for detection because there exists an inconsistency between the classification score and the IoU of a candidate target box when the classifier is trained on source data. More specifically, it is hard to distinguish high quality boxes from low quality ones as a high score does not

necessarily mean a good IoU box, and vice-versa. Also, the pseudo boxes assigned by softmax classifier are not informative enough as these boxes have already been predicted with high probabilities by the same classifier. Instead of using the decision-boundary based classifier, some researchers have proposed to use clustering (Kang et al. 2019; Sener et al. 2016) or label propagation (Zhang et al. 2020) to assign pseudo labels in the domain adaption for classification. Although these assigners have proved their effectiveness, it is impractical to apply them on detection due to the unaffordable memory and computation cost as there are thousands of proposals for each image.

In order to address aforementioned problem, we propose a category dictionary guided (CDG) UDA model to select more reliable and informative pseudo boxes for detection. Specifically, we first learn a dictionary of representative atoms for each category using the source domain features, and use the learned dictionaries to represent each candidate box in the target domain in a collaborative representation manner (Zhang, Yang, and Feng 2011). Then we assign pseudo-labels to these boxes according to their representation residuals w.r.t. each category. One good property of CDG is that the representation residual can not only be used for label assignment but also be used to indicate the quality (e.g., IoU) of the candidate box. By regarding pseudo-boxes as ground truth, we propose a residual weighted self-training pipeline on the target data by assigning different weights to pseudo-boxes based on their representation residuals.

## Related Work

Many advanced computer vision models (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012; Ren et al. 2015; Long, Shelhamer, and Darrell 2015) are based on deep neural networks trained on large scale datasets (Lin et al. 2014; Deng et al. 2009). A domain gap may impair the model's performance on a shifted target data. To address this issue, a variety of domain adaptation methods have been proposed in the fields of classification (Zou et al. 2019; Ganin and Lempitsky 2014; Long et al. 2016), segmentation (Zou et al. 2018; Zhang et al. 2019; Xie et al. 2018; Wu et al. 2018; Murez et al. 2018) as well as detection (Kim et al. 2019b; Chen et al. 2020; He and Zhang 2019; Zheng et al. 2020). Generally speaking, there are three lines of research of UDA in object detection: style transfer (Kim et al. 2019a), feature alignment (Wu et al. 2019; Xu et al. 2020; Li et al. 2020) and self-training (Zou et al. 2018, 2019). In addition, our work is related to dictionary learning for image representation and classification.

**Style Transfer.** Style transfer aims to narrow down the domain gap in pixel-wise features such as color and texture (Inoue et al. 2018). CycleGAN (Zhu et al. 2017) is employed to convert the images in the source domain to domain-transferred images that have similar low level styles to the target data. Style transfer is often combined with other UDA methods for cross-domain object detection.

**Feature Alignment.** Early works in feature alignment mainly utilize adversarial loss to align the visual features (He and Zhang 2019). Later, it was thought that different from classification, object detection focuses more on local areas that contain objects of interest. Based on this observation, researchers attempted to minimize the domain discrepancy at the instance level after the RoIAlign (He et al. 2017) or combine local alignment with global alignment to ensure a stronger adaptation (Saito et al. 2019). Recently, weighted adversarial learning has been proposed to assign different adaptation weights to different areas or instances by assuming that not all samples or regions are equally transferable (Chen et al. 2020). The weighed adaptation can alleviate the negative transfer caused by brute alignment of features in different domains.

The aforementioned feature alignment methods are class-agnostic, i.e., the adapted model by global image-level alignment cannot distinguish the objects belonging to different classes. Following the recent work in classification (Kang et al. 2019), class-specific alignment methods have also been explored in object detection (Zheng et al. 2020). It should be noted that category-specific feature alignment also requires a pseudo-label for each sample.

**Self-training.** Assigning pseudo-labels to target domain samples can help the detector to explore domain specific knowledge directly. Kim *et al.* (Kim et al. 2019a) employed a high threshold to choose reliable pseudo boxes and proposed a weak negative mining operation to reduce the effect of false negatives. In (RoyChowdhury et al. 2019), easy pseudo boxes are obtained from high-confidence predictions by some detectors and hard boxes are selected from a tracker. In (Khodabandeh et al. 2019), a robust self-training scheme was proposed to deal with noisy labels by introducing an auxiliary classifier. Different from the above methods, where the pseudo-boxes are all determined by the category scores predicted by the original softmax based classifier, we propose a category dictionary guided model to generate pseudo labels from the perspective of feature representation, which can not only reflect the class confidence (for classification) but also the quality of the box (for regression).

**Dictionary Learning.** Dictionary learning (DL) aims to learn an effective data representation model from training data, which can be used for classification tasks by exploiting the class label information. Some DL methods (Gu et al. 2014; Mairal et al. 2009) learn a shared dictionary for all classes and the discrimination capability is imposed on the coding coefficients, while some DL methods (Yang et al. 2010; Cai et al. 2016) learn a structured dictionary to promote discrimination between classes. Our model adopts the latter strategy, and we learn a sub-dictionary for each class by minimizing the representation residual.

## Proposed Method

In this work, we address the UDA problem for object detection by a self-training manner. Suppose we have access to a set of source images $x_s$ with labels $y_s$ and box annotations $b_s$. Meanwhile, there is a set of unlabeled target images $x_t$ drawn from a different distribution from $x_s$. Our goal is to learn a detector that can behave well on the target data.

The framework of our proposed CDG-UDA model is shown in Fig. 1. Fig. 1 (a) shows the pipeline of self-training, where images from two domains are fed to Faster RCNN to

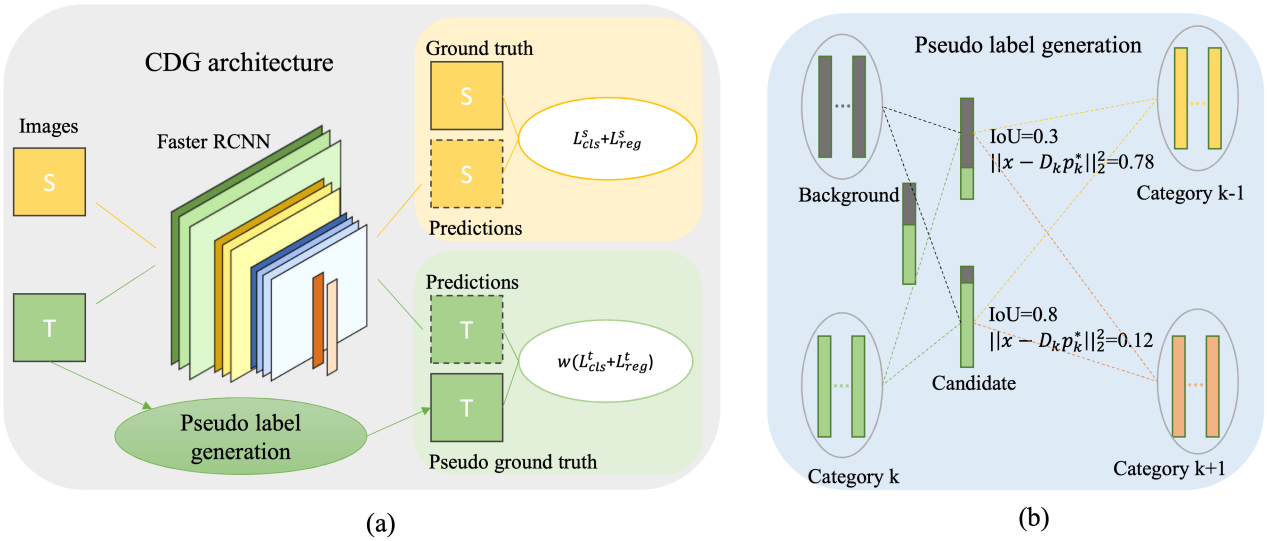


Figure 1: Illustration of the proposed category dictionary guided UDA (CDG-UDA) model for object detection. (a) The architecture of the CDG-UDA. The green parts describe the target data flow while the yellow parts show the source domain data flow. (b) The pseudo label generation and assignment process (best viewed in color). Each ellipse represents the learned dictionary for a category.

learn a shared discriminative detector. Source data are supervised by the ground-truth (GT) while target data are supervised by the weighted pseudo GT boxes generated by the process shown in Fig. 1 (b). In the following, we introduce the major components of our CDG framework in detail.

## Dictionary Learning (DL) for Pattern Classification

Given a feaure matrix $X_k \in R^{l\times n}$ extracted from category $k$ in the source domain, where each column of $X_k$ is the $\ell_2$ normalized feature vector of a sample in category $k$, we learn a dictionary $D_k \in R^{l\times m}$ for category $k$ by solving the following $\ell_1$ sparse optimization problem:

$$\min_{D_k, A_k} (||X_k - D_k A_k||^2_F + \lambda_1 ||A_k||_1) \quad s.t. \quad d_j^T d_j = 1, \tag{1}$$

where $l$ is the feature dimension, $n$ is the total number of samples, $m$ is the number of atoms in $D_k$, $A_k \in R^{m\times n}$ is the sparse coefficient matrix, $\lambda_1$ is a regularization parameter, and each atom $d_j$ in $D_k$ is constrained to have $\ell_2$ unit length. Eq. 1 can be solved by alternative minimization algorithms such as the one in (Kim et al. 2007).

By learning a dictionary for each category, we obtain a whole dictionary $D = [D_1, \ldots, D_K]$ for all the $K$ categories. Given a test sample $x$, we adopt the $\ell_2$-regularized collaborative representation (Zhang, Yang, and Feng 2011) scheme to encode it over $D$:

$$p^* = \text{argmin}_p ||x - Dp||^2_2 + \lambda_2 ||p||^2_2, \tag{2}$$

where $\lambda_2$ is a regularization parameter, $p = [p_1, \ldots, p_K]$ is the coding coefficient vector and $p_k$ is the associated sub-vector for category $k$. Eq. 2 has a closed-form solution (Zhang, Yang, and Feng 2011): $p^* = (D^T D + \lambda_2 I)^{-1} D^T x$. We can then calculate the representation residual of $x$ by each category: $r_k = ||x - D_k p^*_k||^2_2$, and sample $x$ can be classified to the category which has the smallest residual $r_k$.

## Category Dictionary Guided Pseudo-box Generation

The DL described in Section 3.1 is traditionally employed for image classification. In this section, we adapt it to cross-domain object detection and use it for pseudo-box generation. We learn a dictionary $D_k$ for each category $k$ in the source domain so that $D_k$ can approximate each sample belonging to category $k$ as a linear combination of all its atoms. The features we use to learn $D_k$ are extracted from the last fully connected layer in the detection head since it is close to the final classifier. For foreground categories, we consider all the GT boxes as the proposals in the second stage of Faster RCNN and extract their features to learn the dictionary. For the background category, we randomly generate some boxes whose maximum IoU with all GT boxes fall into the interval (0, 0.1]. This enables us to extract background features from the areas close to the GT boxes.

In classification (Kang et al. 2019) and segmentation (Zhang et al. 2019), it has been demonstrated that designing an extra annotator independent of the softmax classifier can select better pseudo labels and thus lead to better performance. Different from these works, where the prototype of each category is simply obtained by calculating the centroid of all sample features, in our method a more representative dictionary is learned for each category, and all dictionaries are used together to encode the candidate box for classification as well as quality estimation.

**Candidate Box Representation.** A detector trained from source data first forwards each target image to generate candidate boxes. The features of these boxes are extacted from the last layer of the detection head. Then we encode the fea-

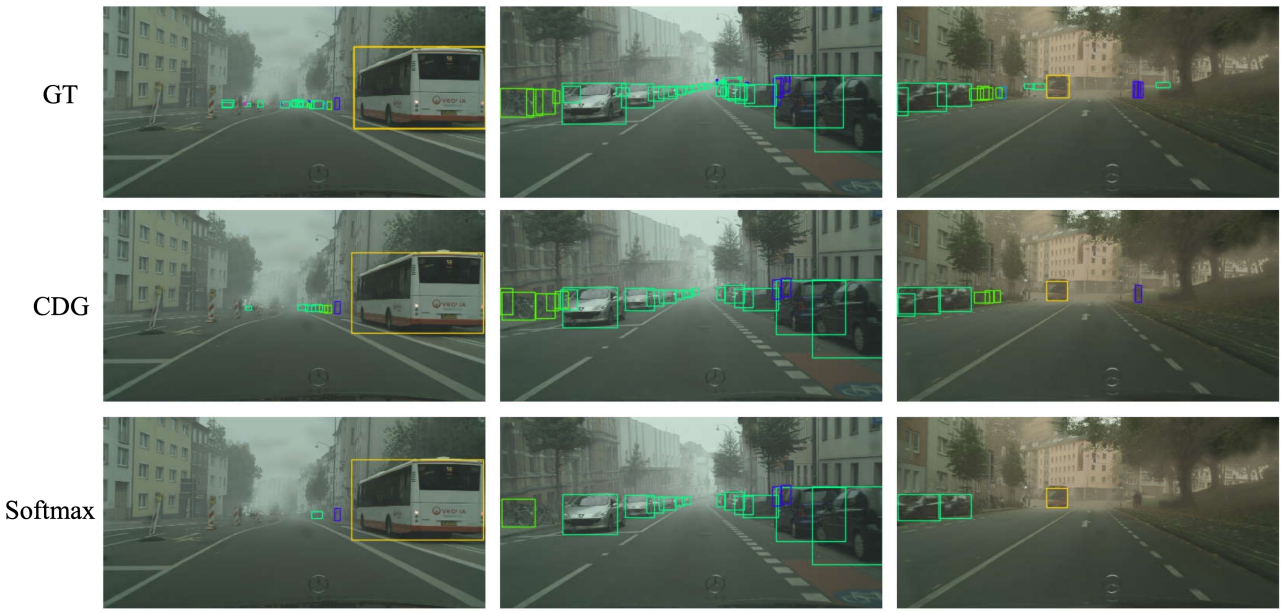


Figure 2: Example pseudo boxes on the foggy cityscapes dataset (Cordts et al. 2016). From top row to bottom row: GT boxes, pseudo boxes selected by CDG and softmax, respectively. Different colored boxes represent different categories.

ture $x$ of each candidate by the dictionary $D$ as in Eq. 2 and calculate the representation residual $r_k$ for each category.

The residual $r_k$ measures how well the sample $x$ can be represented by the $k^{th}$ category. Each candidate box can be seen as a combination of foreground and background, as shown in Fig. 1(b). Since the dictionary for each foreground category is learned by candidate boxes whose IoUs are 1, the feature $x$ of a box that has a higher IoU is likely to have a smaller $r_k$, and vice versa. This means that $r_k$ can roughly indicate how much content in the candidate box belongs to the background and how much belongs to the foreground.

It should be noted that the representation residual is different from the classification score predicted by a decision boundary based classifier trained by manually defined positive samples (IoU larger than 0.5) and negatives samples (IoU smaller than 0.5). The representation based pseudo box selection process with dictionary $D$ is independent of the boundary based softmax classifier and can provide additional information to the detector.

**Pseudo Label Assignment.** For each candidate box $x$ in target image, we have a representation residual vector $r = [r_1, \ldots, r_K]$. We assign the pseudo label to $x$ based on not only the minimal representation residual among $r$, but also a threshold $\theta$ to filter out unreliable candidate boxes. A pseudo GT box is assigned to category $k^*$ when the following two conditions are met:

$$\begin{cases} k^* = \operatorname{argmin}_k r_k \\ r_k - r_{k^*} > \theta, \quad \forall k \neq k^*, \end{cases} \tag{3}$$

The first condition selects the most representative (i.e., smallest residual) category for candidate box $x$, while the second condition ensures that the representation residual $r_{k^*}$ is sufficiently smaller than $r_k$ for other categories. This can remove many uncertain candidate boxes which can be similarly represented by several categories.

Our CDG method can generate more reliable and informative pseudo boxes than the softmax classifier. Here by 'reliable and informative', we mean that a pseudo box is positive but is hard to be detected by the detector. Fig. 2 shows some examples of selected pseudo boxes on the foggy cityscapes (Cordts et al. 2016) dataset. The first row shows the ground truth. The second and third row show the pseudo boxes selected by CDG and softmax classifier, respectively. One can see that CDG can select more positive boxes and assign them with trustable labels, especially for small objects far from the camera, such as 'bike' and 'car'. These boxes are informative for the current detector as it cannot detect them with high confidence.

The classification schemes between dictionary learning and softmax are different. The former classifies one example by considering how well it can be represented by several dictionaries while the latter classifies one example by measuring the distance between the sample and the decision boundary. There may exist some boxes (near the decision boundary) that are ambiguous for the original classifier but can well be classified by the learned dictionary. These boxes are difficult for current detector which uses softmax and can provide extra knowledge for the detector to learn.

## Residual Weighted Self-training

As we discussed in the section of candidate box representation, the representation residual $r_{k^*}$ can indicate how much content in the candidate box belongs to background and foreground. In other words, $r_{k^*}$ can reflect the quality or IoU of a candidate box. We calculate the IoUs of all candidate boxes in the target dataset (the Watercolor dataset (Kim

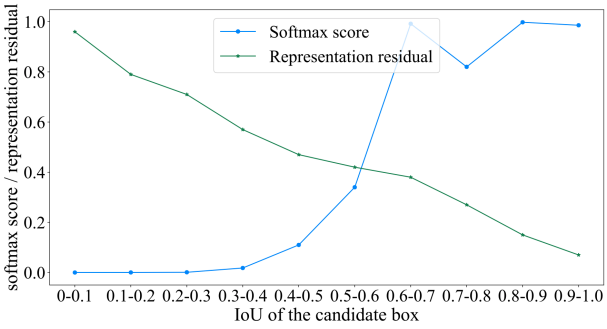


Figure 3: The softmax classification scores (blue) and representation residuals by CDG (green) vs. IoU of the candidate boxes. The results are calculated on the Watercolor (Kim et al. 2019a) dataset.

et al. 2019a) is used), and plot their relationships with the scores predicted by softmax and representation residuals by CDG in Fig. 3. We can see that there is indeed a nearly linear correlation between $r_{k^*}$ and IoU. In contrast, the softmax scores can be either over-confident or under-confident for different IoUs.

The observations in Fig. 3 motivate us to make use of the selected pseudo boxes for model training. A high quality pseudo box (smaller $r_{k^*}$) should contribute more to the loss, and vice versa. So, we propose a residual weighted self-training scheme, whose loss function is defined as:

$$L = \mu\left(L^s_{cls} + L^s_{reg}\right) + \sum_{j=1}^{M} w_j \left(L^t_{cls}\left(I, y^t_j\right) + L^t_{reg}\left(I, b^t_j\right)\right), \tag{4}$$

where $L^s_{cls}$ and $L^s_{reg}$ are the classification and regression losses in source domain, and $L^t_{cls}$ and $L^t_{reg}$ are those in target domain; $\mu$ is a parameter to balance source and target losses; $M$ is the number of assigned pseudo-boxes for the input image $I$; $y^t_j$ is the label of pseudo GT box $b^t_j$; and $w_j$ is the training weight assigned to pseudo GT box $b^t_j$, and it is designed as follows:

$$w_j = 1 - \frac{s}{r_{\max} - r_{\min}}\left(r_j - r_{\min}\right), \tag{5}$$

where $s \geq 0$ is a hyper-parameter, $r_j$ is the representation residual for $b^t_j$ (determined by Eq. 3), and $r_{max}$ and $r_{min}$ are the maximum and minimum representation residuals among all the pseudo boxes in the target data. The function in Eq. 5 ensures that $w_j$ increases linearly with the decrease of $r_j$, i.e., a higher quality pseudo box will have a higher weight.

# Experiments

## Experiments Setup

Following the common self-training pipeline (Zhang et al. 2019), an initialized model is first utilized to generate pseudo-labels by CDG and then the detector is re-trained on both source data and target data by regarding pseudo-labels as ground truth using a standard training pipeline for detection. In order to get higher quality of pseudo labels, the initialized model for multi-class datasets is trained from both source data and target data by a global feature alignment method (Tzeng et al. 2017). We utilize a stage-wise training scheme where self-training is performed for several stages and in each stage the pseudo-labels are re-generated by the model in the last stage.

We evaluate our CDG-UDA model on three benchmarks, including Sim10k (Johnson-Roberson et al. 2016) to Citycapes (Cordts et al. 2016), Cityscapes to Foggy Cityscapes (Cordts et al. 2016) and Pascal VOC (Everingham et al. 2010) to Watercolor (Inoue et al. 2018), to demonstrate its effectiveness on adapting both dissimilar and similar domains. Our model is based on one most typical detector Faster RCNN (Ren et al. 2015) with RoIAlign (He et al. 2017). The shorter side of the input image is resized to 600. During the stage-wise training, we set the initial learning rate as 0.001 and train the model for 20 epochs. The learning rate is decayed by a factor 10 at the $10th$ epoch and $15th$ epoch, respectively. Batch size is set as 1 and SGD is used as the optimizer. For evaluation metric, we report the mean average precision (mAP) at threshold 0.5.

For the parameters associated with DL, we just simply follow the common settings (Zhang, Yang, and Feng 2011) to set $\lambda_1$, $\lambda_2$ as 0.01 and 0.1. For $m$, we choose 64 for Watercolor and 192 for other datasets. We find that the model performance is robust to the value of $m$. Then we mainly have three parameters $\mu$, $s$ and $\theta$. We set $\mu$ to 2 for all datasets. For $s$, we set it to 0.4 and 0.5 for Watercolor and other datasets, respectively. As $\theta$ controls the number of selected pseudo-labels, it varies among different datasets. Specifically, we set it as 0.1 on Sim10k and Foggy Cityscapes. Due to the fact that Watercolor has a severe class imbalance problem ('person' occupies the majority of dataset), the initialized classifiers for different categories are also imbalanced. Given this observation, we set $\theta$ to 0.2 for 'person' and a higher value 0.5 for all the other categories.

## Comparison with State-of-the-arts

**Multi-class Normal to Foggy Adaptation.** In this task, Citycapes (Cordts et al. 2016) servers as the source domain, which contains of 2,975 images from 8 classes. The Foggy Cityscapes (Cordts et al. 2016) is selected as the target domain, which has 2,975 images for training and 500 images for testing. On this task, the majority of the competitive methods are based on feature alignment. Note that feature alignment on multi-class datasets is very challenging because of the imbalance of classes. Class-agnostic feature alignment is hard to guarantee that samples from different categories in the target domain can be properly separated.

MTOR (Cai et al. 2019) treats UDA for detection as a semi-supervised problem and enforces consistency regularization on object relations. NL (Khodabandeh et al. 2019) addresses domain adaptation from the perspective of robust learning from noisy labels. MA (He and Zhang 2019) performs feature alignment on multiple levels of feature maps. SW (Saito et al. 2019) and HTCN (Chen et al. 2020) assign different weights to hardly transferrable and easily transferrable features. GPA (Xu et al. 2020) applies category-wise feature adaptation.

The results by competing methods are showed in Table 1.

| Method | B | Bus | Bicycle | Car | Motorcycle | Person | Rider | Train | Truck | mAP |
|---|---|---|---|---|---|---|---|---|---|---|
| NL (Khodabandeh et al. 2019) | V | 35.1 | 42.15 | 49.17 | 30.07 | 45.25 | 26.97 | 26.85 | 36.03 | 36.45 |
| MTOR (Cai et al. 2019) | R | 30.6 | 41.4 | 44 | 21.9 | 38.6 | 40.6 | 28.3 | 35.6 | 35.1 |
| SW (Saito et al. 2019) | V | 36.2 | 35.3 | 43.5 | 30 | 29.9 | 42.3 | 32.6 | 24.5 | 34.3 |
| MA (He and Zhang 2019) | R | 28.2 | 39.5 | 43.9 | 23.8 | 39.9 | 33.3 | 29.2 | 33.9 | 34 |
| CF (Zheng et al. 2020) | V | 43.2 | 37.4 | 52.1 | 34.7 | 34 | 46.9 | 29.9 | 30.8 | 38.6 |
| HTCN (Chen et al. 2020) | V | 33.2 | 47.5 | 47.9 | 31.6 | 47.4 | 40.9 | 32.3 | 37.1 | 39.8 |
| PD (Wu et al. 2019) | V | 33.12 | 43.41 | 49.63 | 21.98 | 45.75 | 32.04 | 29.59 | 37.08 | 36.57 |
| GPA (Xu et al. 2020) | R | 32.9 | 46.7 | 54.1 | 24.7 | 45.7 | 41.1 | 32.4 | 38.7 | 39.5 |
| Ours | V | 47.5 | 38.9 | 53.1 | 38.3 | 38 | 47.4 | 41.1 | 34.2 | 42.3 |

Table 1: Results of different methods on 'Cityscape to Foggy'.'B,V,R' means 'Backbone','VGG' and 'Resnet', respectively.

| Method | Bike | Bird | Car | Cat | Dog | Person | mAP |
|---|---|---|---|---|---|---|---|
| SW (Saito et al. 2019) | 82.3 | 55.9 | 46.5 | 32.7 | 35.5 | 66.7 | 53.3 |
| DM (Kim et al. 2019c) | - | - | - | - | - | - | 52 |
| WST+BSR (Kim et al. 2019a) | 75.6 | 45.8 | 49.3 | 34.1 | 30.1 | 64.1 | 49.9 |
| PD (Wu et al. 2019) | 95.8 | 54.3 | 48.3 | 42.4 | 35.1 | 65.8 | 56.9 |
| Ours | 97.7 | 53.1 | 52.1 | 47.3 | 38.7 | 68.9 | 59.7 |

Table 2: Results of different methods on 'VOC to Watercolor'

| Method | B | Source only | mAP |
|---|---|---|---|
| NL (Khodabandeh et al. 2019) | V | 31.08 | 42.6 |
| MTOR (Cai et al. 2019) | R | 39.4 | 46.6 |
| SW (Saito et al. 2019) | V | 34.6 | 42.3 |
| MA (He and Zhang 2019) | R | 30.1 | 41.2 |
| CF (Zheng et al. 2020) | V | 35 | 43.8 |
| HTCN (Chen et al. 2020) | V | 34.6 | 42.5 |
| Ours | V | 34.9 | 48.8 |

Table 3: Results of different methods on 'Sim10k to Cityscapes'. 'B' means backbone.

For the employed backbone network,"V" means VGG16 and "R" means ResNet101. Our CDG method with VGG16 backbone achieves an mAP of 42.3, outperforming the second best model HTCN by 2.5 points. It is worth mentioning that the recent feature alignment methods for cross-domain object detection is becoming more and more complex in order to address the various challenges discussed in the introduction section. In contrast, our model is much simpler and it achieves the leading performance.

**Realistic to Artistic Adaptation.** In this experiment, we use Pascal VOC2007 trainval and VOC2012 trainval as the source domain, and Watercolor as the target domain. Pascal VOC (Everingham et al. 2010) is a real world image dataset which consists of a total of 16,551 images from 20 categories. Watercolor (Inoue et al. 2018) is an artistic dataset that has 6 common classes with VOC. It has 1,000 images for training and another 1,000 images for testing. For fair comparison, we follow the settings in competing papers (Wu et al. 2019; Saito et al. 2019; Kim et al. 2019c) and use ResNet101 as our backbone.

The results on this task are shown in Table 2. The Watercolor dataset has serious class imbalance, where categories like 'car', 'cat' and 'dog' have much fewer images than other categories. In this case, perfect category-wise feature alignment is very hard to achieve. However, our proposed CDG can significantly improve the performance of these rare categories, as can be seen in Table 2. Overall, CDG achieves an mAP of 59.7, outperforming the second best method PD (Wu et al. 2019) by 2.8 points.

**Synthetic to Real Adaptation.** We then evaluate CDG's adaptation performance from synthetic images to real images. Sim10k is used as the source dataset, which comprises 10,000 synthetic images collected from the computer game Grand Theft Auto (GTA). All these images are used for training. Cityscape is used as the target dataset, which contains 2,975 training images and 500 test images captured from real city streets. Following previous works (Chen et al. 2020; Zheng et al. 2020), in this task we only focus on the category 'car'.

The results are shown in Table 3. It can be seen that our model outperforms the comparison methods using the same VGG16 backbone by a large margin. It also performs better than MTOR (Chen et al. 2020) which utilizes ResNet101 as the backbone. This demonstrates that exploiting the target domain knowledge by the pseudo labels in a self-training manner can achieve better domain adaptation results.

## Ablation Study

**Component Analysis.** In this section, we conduct ablation studies to investigate the effect of some elements in our CDG-UDA method on the final performance. The 'VOC to Watercolor' setting is used and the results of different variants of CDG-UDA are shown in Table 4. Here, 'Source only' means that the model is trained only on source data.

The method 'Softmax' uses softmax classifier as the an-

| Method | Bike | Bird | Car | Cat | Dog | Person | mAP |
|---|---|---|---|---|---|---|---|
| Source only | 86.6 | 39.4 | 37 | 23.2 | 20.7 | 56.2 | 43.9 |
| Softmax | 89.8 | 52.3 | 53 | 40.1 | 31.6 | 65.3 | 55.3 |
| CDG-stage 1 | 95.9 | 51.7 | 53 | 41.2 | 40.8 | 65.8 | 58 |
| CDG-weighted-stage 1 | 97 | 53.7 | 51.1 | 43.5 | 39.7 | 66 | 58.5 |
| CDG-stage 2 | 99.9 | 51.1 | 52.3 | 42.9 | 40.3 | 69 | 59.3 |
| CDG-weighted-stage 2 | 97.7 | 53.1 | 52.1 | 47.3 | 38.7 | 68.9 | 59.7 |

Table 4: Result of ablation study on 'VOC to Watercolor'

notator for pseudo box generation, while the method 'CDG-stage 1' applies our CDG based self-training (all weights $w_j$ are fixed to 1) for only one stage. One can see that 'CDG-stage 1' outperforms 'softmax' by 2.7 points, which demonstrates that the proposed CDG is more effective than the softmax based classifier. When we apply self-training for two stages, the method 'CDG-stage 2' improves the mAP over 'CDG-stage 1' by 1.3 points. We found that further increasing the training stages will not bring extra improvement.

By assigning different weights to the pseudo boxes based on their representation residuals, method 'CDG-weighted-stage 1' improves the performance of 'CDG-stage 1' by 0.5 point, while 'CDG-weighted-stage 2' (i.e., the default CDG method) improves 'CDG-stage 2' by 0.4 point. This validates that the weighed self-training can consistently bring extra performance gains at each training stage.

**Hyper-parameters Sensitivity.** In this section, we give a more detailed ablation study for three hyper-parameters in the residual weighted self-training including $s$, $\mu$ and $\theta$.

$s$ determines the weight assignment function for each pseudo-box in the self-training process. A large $s$ means pseudo-boxes with large residuals will be assigned to smaller weights. If $s$ is 0, all weights $w_j$ are fixed to 1. Experiments on Watercolor in the first stage are conducted to investigate its influence on the model's performance by setting different values for $s$. The results are shown in Table 5. It can seen that a high or low value of $s$ can degrade the performance and the peak accuracy is achieved when $s$ is 0.4.

| $s$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
|---|---|---|---|---|---|---|---|
| mAP | 58.0 | 58.1 | 58.1 | 58.3 | 58.5 | 58.3 | 58.2 |

Table 5: Sensitivity of $s$ on 'VOC to Watercolor'.

In the residual weighted self-training, $\mu$ is a parameter to balance the loss between the source data and target data. We set different values for $\mu$ in the first training stage on Watercolor dataset. The results are shown in Table 6. We can see that the setting of $\mu$ can significantly influence the accuracy of the final detector. If the model is only trained on the pseudo-labeled target data ($\mu = 0$), the mAP is not high, which is because the error labels may mislead the detector training. When the model is trained on both domains ($\mu > 0$), the source data can provide more precious supervision signals preventing the detector from being misguided by the pseudo-labeled target data. However, a larger $\mu$ can make the gradients dominated by the source data as a result of which, the detector cannot learn domain specific knowledge from the target domain.

| $\mu$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| mAP | 55.9 | 57.0 | 58.0 | 55.9 |

Table 6: Sensitivity of $\mu$ on 'VOC to Watercolor'.

$\theta$ controls the number of pseudo labels selected from the target data for self-training. A small $\theta$ can select more pseudo boxes with more noises while a relatively large $\theta$ can select fewer pseudo boxes with fewer noises. We provide the sensitive analysis for $\theta$ on 'Sim10k to Cityscape' in Table 7. We can see that the mode achieves its highest performance when $\theta$ is 0.1. We also observe that Watercolor has a different optimal $\theta$ (0.2 for 'person' and 0.5 for others) with other two datasets, the reason of which is that some classes in Watercolor have very few instances (no more than 50) for some classes like 'bird', 'cat' and 'dog' that a small $\theta$ will bring too many false positives.

| $\theta$ | 0 | 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|
| mAP | 48.5 | 48.8 | 46.3 | 45.7 |

Table 7: Sensitivity of $\mu$ on 'Sim10k to Foggy Cityscapes'.

## Conclusion

In this paper, we proposed a novel category dictionary guided (CDG) unsupervised domain adaptation (UDA) model for cross-domain object detection. The CDG-UDA model explores the domain specific knowledge by learning category-specific dictionaries to generate more reliable and informative pseudo boxes. Meanwhile, the representation residuals by the dictionaries are also good indicators of pseudo box qualities. A residual weighted stage-wise self-training scheme was consequently proposed to alleviate the effect of low-quality labels and train a more robust cross-domain detector. The experiments on three different datasets demonstrated the superiority of our proposed CDG model to the popular feature alignment models. Our method indicated a new and promising direction for future research of cross-domain object detection.

## References


Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Cai, S.; Zhang, L.; Zuo, W.; and Feng, X. 2016. A probabilistic collaborative representation based approach for pattern classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2950–2959.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing Transferability and Discriminability for Adapting Object Detectors. *arXiv preprint arXiv:2003.06297* .

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88(2): p.303–338.

Ganin, Y.; and Lempitsky, V. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* .

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *Advances in neural information processing systems*, 793–801.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z.; and Zhang, L. 2019. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 6668–6677.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5001–5009.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2016. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983* .

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Khodabandeh, M.; Vahdat, A.; Ranjbar, M.; and Macready, W. G. 2019. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 480–490.

Kim, S.; Choi, J.; Kim, T.; and Kim, C. 2019a. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 6092–6101.

Kim, S.-J.; Koh, K.; Lustig, M.; Boyd, S.; and Gorinevsky, D. 2007. A method for large-scale l1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing* 1(4): 606–617.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019b. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12456–12465.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019c. Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Object Detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* doi:10.1109/cvpr.2019.01274. URL http://dx.doi.org/10.1109/CVPR.2019.01274.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lee, S.; Kim, D.; Kim, N.; and Jeong, S.-G. 2019. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 91–100.

Li, W.; Li, F.; Luo, Y.; and Wang, P. 2020. Deep Domain Adaptive Object Detection: a Survey. *arXiv preprint arXiv:2002.06797* .

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, 136–144.

Mairal, J.; Ponce, J.; Sapiro, G.; Zisserman, A.; and Bach, F. R. 2009. Supervised dictionary learning. In *Advances in neural information processing systems*, 1033–1040.

Murez, Z.; Kolouri, S.; Kriegman, D.; Ramamoorthi, R.; and Kim, K. 2018. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4500–4509.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

RoyChowdhury, A.; Chakrabarty, P.; Singh, A.; Jin, S.; Jiang, H.; Cao, L.; and Learned-Miller, E. 2019. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 780–790.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6956–6965.

Sener, O.; Song, H. O.; Saxena, A.; and Savarese, S. 2016. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, 2110–2118.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Wu, A.; Han, Y.; Zhu, L.; and Yang, Y. 2019. Instance-Invariant Adaptive Object Detection via Progressive Disentanglement. *arXiv preprint arXiv:1911.08712* .

Wu, Z.; Han, X.; Lin, Y.-L.; Gokhan Uzunbas, M.; Goldstein, T.; Nam Lim, S.; and Davis, L. S. 2018. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 518–534.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, 5423–5432.

Xu, M.; Wang, H.; Ni, B.; Tian, Q.; and Zhang, W. 2020. Cross-domain Detection via Graph-induced Prototype Alignment. *arXiv preprint arXiv:2003.12849* .

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1426–1435.

Yang, M.; Zhang, L.; Yang, J.; and Zhang, D. 2010. Metaface learning for sparse representation based face recognition. In *2010 IEEE International Conference on Image Processing*, 1601–1604. IEEE.

Zhang, L.; Yang, M.; and Feng, X. 2011. Sparse representation or collaborative representation: Which helps face recognition? In *2011 International conference on computer vision*, 471–478. IEEE.

Zhang, Q.; Zhang, J.; Liu, W.; and Tao, D. 2019. Category Anchor-Guided Unsupervised Domain Adaptation for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, 433–443.

Zhang, Y.; Deng, B.; Jia, K.; and Zhang, L. 2020. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, 781–797. Springer.

Zheng, Y.; Huang, D.; Liu, S.; and Wang, Y. 2020. Cross-domain Object Detection through Coarse-to-Fine Feature Adaptation. *arXiv preprint arXiv:2003.10275* .

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 687–696.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, 5982–5991.

Zou, Y.; Yu, Z.; Vijaya Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.