Domain Contrast for Domain Adaptive Object Detection

Feng Liu, Xiaosong Zhang, *Student Member, IEEE*, Fang Wan, *Student Member, IEEE*, Xiangyang Ji, *Member, IEEE*, and Qixiang Ye, *Senior Member, IEEE*

***Abstract*—Despite of the substantial progress of visual object detection, models trained in one video domain often fail to generalize well to others due to the change of camera configurations, lighting conditions, and object person views. In this paper, we present Domain Contrast (DC), a simple yet effective approach inspired by contrastive learning for training domain adaptive detectors. DC is deduced from the error bound minimization perspective of a transferred model, and is implemented with cross-domain contrast loss which is plug-and-play. By minimizing cross-domain contrast loss, DC transfers detectors across domains while naturally alleviating the class imbalance issue in the target domain. DC can be applied at either image level or region level, consistently improving detectors' discriminability while maintaining the transferability. Extensive experiments on commonly used benchmarks show that DC improves the baseline and state-of-the-art by significant margins, while demonstrating great potential for large domain divergence. Code is released at github.com/PhoneSix/Domain-Contrast.**

***Index Terms*—Domain Adaptation, Visual Object Detection, Contrastive Learning, Transfer Learning.**

I. Introduction

MODERN object detectors [1], [2] have achieved unprecedented progress with the rise of convolutional neural networks (CNNs). However, their practical application to real-world video scenarios remains limited for the following two reasons: 1) Supervised learning of detectors for different scenarios requires repeated human effort on data annotation, and 2) offline-trained detectors typically degrade with changes in the scene or camera. Domain adaptive detection, which transfers detectors trained within a label-rich domain ($i.e.$, annotated datasets) to an unlabeled domain ($i.e.$, real-world video scenarios), have attracted increasing attention because of their potential to solve these problems [3], [4], [5], [6].

Domain adaptive detection is usually explored with unsupervised domain adaptation (UDA) methods. One line of UDA methods utilizes adversarial generative models as a style transformer to convert images (with identity annotations) of a source domain into a target domain [7], [8]. The other line of methods attempts to match the feature distributions of source and target domains to maintain model performance [9], [10]. This line of methods typically maximizes the "transferability" of models by aligning the feature distributions of domains [3]. The underlying hypothesis is that accurate feature alignment across domains produces good transferability.

Recent studies [11] have shown that transferability and discriminability are in fact two sides of a same coin. The transferring process, $e.g.$, minimizing Maximum Mean Discrepancy (MMD) [12], could deteriorate the discriminability of models and features [13]. This would be more severe in the object detection problem, considering the large imbalance of negative and positive instances which need to be classified. Because of the large amount of negative instances from the backgrounds, slight degradation of model discriminability could cause a significant increase of false positives in the target domain. There is a requirement to exploit domain adaptive methods which can comprise both the transferability and discriminability of features and detectors.

In this paper, we propose a novel Domain Contrast (DC) approach for domain adaptive object detection, with the aim to maximize model discrimination capacity in the target domain by alternating discriminative learning and domain transfer. We derive the DC method from the perspective of error bound minimization when transferring detectors from a source to a target domain. Minimizing error bound is converted to optimize DC loss, which defines the cross-domain similarity and inter-class distance for each min-batch of samples. DC guarantees the discriminability of transferred detectors by minimizing the similarity between samples from different categories in a mini-batch, Fig. 1. DC preserves the transferability of detectors by maximizing the $cosine$ similarity between each sample with its cross-domain counterparts.

To define DC, training sample images are translated from the source/target to the target/source domain via a CycleGAN method [14]. Other commonly used stye-transfer methods can be also combined with the proposed DC method in a plug-and-play fashion. With translated images, a detector is trained using annotated samples in the source domain by minimizing the object detection loss. The detector is then fine-tuned by optimizing the DC loss which targets at minimizing the domain divergence and maximizing the transferability between the source and the target domain. In turn, the detector is trained using the DC loss from the target to the source domain. With the simple yet effective DC learning, the features are adapted to both the target and source domain data, and have stronger representative capacity. DC loss is applied at both image-level and region-level, consistently improving the transferability and discriminability of detectors across domains for higher

Feng Liu is with the School of Microelectronics, University of Chinese Academy of Sciences. Xiaosong Zhang and Qixiang Ye are with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences. Fang Wan is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China, 100049. Emails: liufeng20@mails.ucas.ac.cn, zhangxiaosong18@mails.ucas.ac.cn, wanfang@ucas.ac.cn, qxye@ucas.ac.cn. Xiangyang Ji is with the Department of Automation, Tsinghua University. Qixiang Ye is the corresponding author.

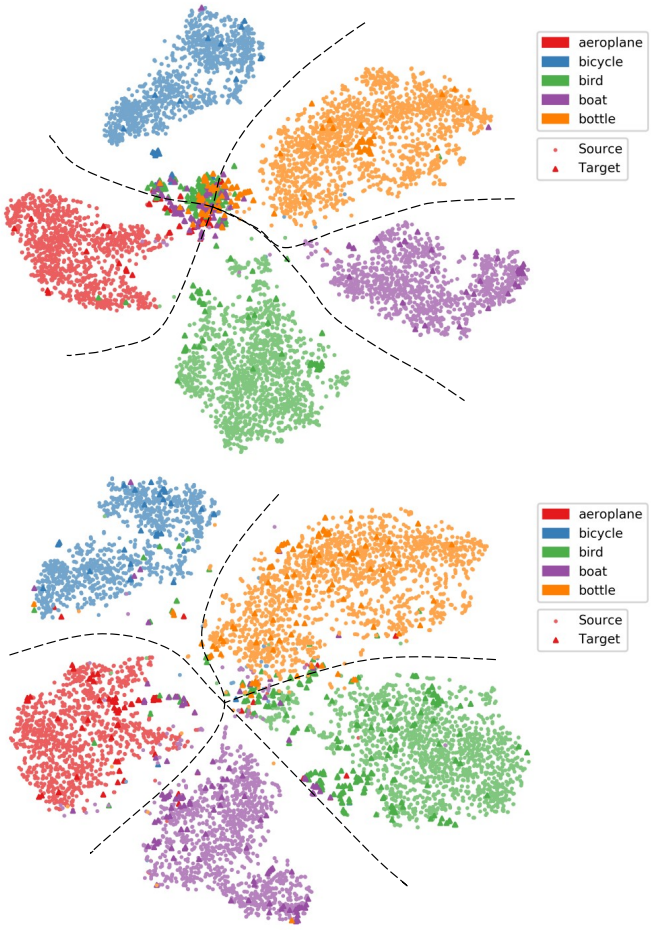


Fig. 1: Samples distribution before (upper) and after (lower) domain adaptation with Domain Contrast (DC). DC simultaneously improves model discriminability, $i.e.$, maximizing the distances between samples from different categories, and aligns samples of same categories from different domain. Data points are features from PASCAL VOC 07 dataset and figure is drawn with t-SNE. (Best viewed in color)

detection performance.

The contributions of this work are summarized as follows:

- A simple-yet-effective domain contrast (DC) approach, which improves model discriminability in the target domain while preserving feature alignment ($i.e.$, transferability) between the source domain and the target domain. The plausibility of DC loss for domain adaptation is justified from a perspective of error bound minimization.
- A domain adaptive detector considering the class imbalance issue. The detector leverages DC loss at image level and region level to handle the detector transfer problem.
- State-of-the-art detection performance on benchmarks with large domain divergence upon negligible computational cost and complexity of the proposed DC approach.

II. Related Work

While object detection has been extensively investigated from various perspectives, we mainly review the domain adaptive object detection approaches, which are largely driven by unsupervised domain adaptation (UDA) methods. We also reviewed recent advance about constrative learning, which is commonly used to enforce feature learning in a self-supervised manner or minimize the divergence between two domains.

A. Unsupervised Domain Adaptation (UDA)

UDA aims to minimize the classification error when transferring the model trained within a source domain to an unlabelled target domain. UDA has been extensively explored in a number of computer vision specializations, including image classification [15], [16], [17], [18], [19], image recognition [20], [21], and object detection [3], [4], [5], [22].

One line of UDA methods devoted to align feature distributions of source and target domains by minimizing the domain divergence [19], [18]. For example, Maximum Mean Discrepancy (MMD) [12] was proposed as a distance metric to minimize the domain divergence in the Reproducing Kernel Hilbert Space (RKHS) [19], [23].

Progressive domain distance minimization [24] was achieved by mining pseudo labels to fine-tune the model. The descrepancy-based method [25] aligned distributions of source and target domains by learning features to minimize classifiers' output discrepancy. TA^{3}N [26] simultaneously learned and matched temporal dynamics with domain discrepancy for domain alignment. The other line of methods [27], [28] attempted to reduce the domain divergence by taking advantages of adversarial and generative models which confuse domains while aligning feature distributions. The CyCADA method [28] transferred samples across domains at both pixel- and feature-levels. Domain confusion loss [18] was designed to learn domain-invariant features. SAVA [29] minimized domain adversarial loss on discriminative clips for cross-domain video representation alignment. TCoN [30] leveraged an adversarial co-attention mechanism to match the distributions of temporal features between source and target domains. In addition to adversarial alignment, MM-SADA [31] leveraged the correspondence of modalities as a self-supervised alignment approach to reduce domain shift.

B. Contrastive Learning

Contrastive loss [32] aimed to learn a good representation by minimizing the distances between positive pairs and maximizing the distances between negative pairs. Instance Discrimination [33] abandoned the euclidean distance, using a non-parametric softmax classifier. To reduce the computation cost, it adopted noise-contrastive estimation [34] strategy. CPC [35] proposed a probabilistic contrastive loss, InfoNCE loss, to maximize a lower bound on mutual information. Deep InfoMax [36] aggregated local and global mutual information and prior matching together to learn a representation that can perform well on various tasks. CPC [35] and Deep InfoMax [36] were further extended in [37] and [38], respectively. These mutual information-based methods learned from two views. Instance Discrimination [33] learned from two crops of a same image and CPC [35] learned from past and future. Deep InfoMax [36] learned from input and output of the neural network.

CMC [39] designed a loss to enable the network to maximize mutual information among multiple views of the same scene. Moco [40] viewed contrastive learning as dictionary look-up and built a dynamic dictionary with a queue and a moving-averaged encoder. SimCLR [41] conducted data augmentations on images to obtain positive and negative sample pairs. It then applied contrastive loss on them to learn invariant features in a self-supervised way. To address the misaligned issue in the domain adaptive classification task, CAN [42] introduced contrastive loss to perform class-aware alignment, optimizing intra-class and inter-class domain discrepancy.

The main differences between our method and above methods can be conlcuded as two points. First, we derive the contrastive loss from the perspective of error bound minimization, providing a new perspective for understanding contrastive loss. Second, we specifity the image- and instance-level constristive loss for the object detection task, which requires to handle a large amount of instances in each image.

C. Domain Adaptive Detection

Early studies largely followed domain adaptive classification to align the features of source and target domains. DA-Faster R-CNN [3] pioneered these works by minimizing the discrepancy among two domains by exploring both image- and instance-level domain classifier in an adversarial manner. Mean Teacher with object relations [43] was applied for object distribution alignment, while integrating object relations with the measure of consistency cost between teacher and student modules. The Diversify and Match (DM) approach [22] generated various distinctive shifted domains from the source domain and aligned the distribution of the labeled data and encouraged features to be indistinguishable among the domains. Strong-and-Weak [4] method pursue weak alignment of image-level features and strong alignment of region-level features.

Despite progress, the essential difference between domain adaptive detection with domain adaptive classification is unfortunately ignored. On the one hand, source and target domains have distinct scene layouts and object combinations. Therefore, aligning the entire distributions of source and target images is implausible. On the other hand, object detectors face the serious class imbalance issue. Preserving model discriminability during domain adaptive detection is more important than that in image classification [5].

To preserve the discriminability, the Selective Cross-Domain approach [5] attempted aligning discriminative regions, namely those that are directly related to detection. The Hierarchical Transferability Calibration Network harmonized transferability and discriminability for cross-domain detection [11]. Nevertheless, these approaches used complex adversarial training and/or sample interpolation which hinders deployment and therefore practicability.

III. The Proposed Approach

Under the context of unsupervised domain adaptation (UDA) for object detection, we have a fully labeled source domain and a unlabeled domain. Let $\mathcal{S}$ and $\mathcal{T}$ respectively denote a source and a target domain. The corresponding samples of $\mathcal{S}$ and $\mathcal{T}$ are denoted as $\{x_s^i\}_{i=1}^N$ and $\{x_t^i\}_{i=1}^N$. $f_\mathcal{S} : \mathcal{X} \to \{0,1\}$ and $f_\mathcal{T} : \mathcal{X} \to \{0,1\}$ denote functions which map the input samples $\mathcal{X}$ to a binary label space. $f_\mathcal{T}(x_s^i) = f_\mathcal{T}(x_t^i)$ and $f_\mathcal{S}(x_s^i) = f_\mathcal{S}(x_t^i)$ mean that the sample labels are consistent regardless of the domains. Domain adaptation targets at transferring a model ($i.e.$, the detector) optimized for $f_\mathcal{S}$ to $\mathcal{T}$ towards optimizing $f_\mathcal{T}$.

In what follows, we first derive the domain contrast method and domain contrast loss from a perspective of error bound minimization. We then describe the object detector based on domain contrast loss.

A. Domain Contrast

Error Bound Minimization. The source and target domains share an identical label space, but violate the $i.i.d.$ assumption as they are sampled from different data distributions. Given a model hypothesis $h \in \mathcal{H}$, the expected error [44] within the target domain are bounded as

$$\mathcal{R}_\mathcal{T}(h, f_\mathcal{T}) \le \mathcal{R}_\mathcal{T}(h, f_\mathcal{S}) + |\mathcal{R}_\mathcal{T}(h, f_\mathcal{T}) - \mathcal{R}_\mathcal{T}(h, f_\mathcal{S})|, \quad (1)$$

where $\mathcal{R}_\mathcal{T}(h, f_\mathcal{T})$ and $\mathcal{R}_\mathcal{T}(h, f_\mathcal{S})$ respectively denote the empirical error of hypothesis h in the target and source domains. To minimize the error bound defined by Eq. 1 is to minimize $|\mathcal{R}_\mathcal{T}(h, f_\mathcal{T}) - \mathcal{R}_\mathcal{T}(h, f_\mathcal{S})|$ and $\mathcal{R}_\mathcal{T}(h, f_\mathcal{S})$, which aligns the two domains while preserving the discriminability of the trained model in the target domain.

In the context of CNN, with the binary cross entropy loss, we have

$$\begin{aligned}\mathcal{R}_\mathcal{T}(h, f_\mathcal{S}) = \frac{1}{N}\sum_i \Big(& - f_\mathcal{S}(x_t^i)\log(h(x_t^i)) \\ & - (1 - f_\mathcal{S}(x_t^i))\log(1 - h(x_t^i))\Big),\end{aligned} \quad (2)$$

and

$$\begin{aligned}\mathcal{R}_\mathcal{T}(h, f_\mathcal{T}) = \frac{1}{N}\sum_i \Big(& - f_\mathcal{T}(x_t^i)\log(h(x_t^i)) \\ & - (1 - f_\mathcal{T}(x_t^i))\log(1 - h(x_t^i))\Big),\end{aligned} \quad (3)$$

where N denotes the number of samples. Subtracting Eq. 3 from Eq. 2, we have

$$\begin{aligned}&|\mathcal{R}_\mathcal{T}(h, f_\mathcal{T}) - \mathcal{R}_\mathcal{T}(h, f_\mathcal{S})| \\ &\quad = \left|\frac{1}{N}\sum_i \Big(- f'(x_t^i)\log(\frac{h(x_t^i)}{1 - h(x_t^i)})\Big)\right|,\end{aligned} \quad (4)$$

where $f'(x) = f_\mathcal{T}(x) - f_\mathcal{S}(x)$.

For the optimal hypothesis h^* in the source domain, it is assumed that the empirical error in the source domain is small enough, $i.e.$, $\mathcal{R}_\mathcal{S}(h^*, f_\mathcal{S}) \to 0$, and the discriminability of h^* for each sample in $\mathcal{T}$ is smaller than that in S, as

$$|\log(\frac{h^*(x_t^i)}{1 - h^*(x_t^i)})| \le |\log(\frac{h^*(x_s^i)}{1 - h^*(x_s^i)})|. \quad (5)$$

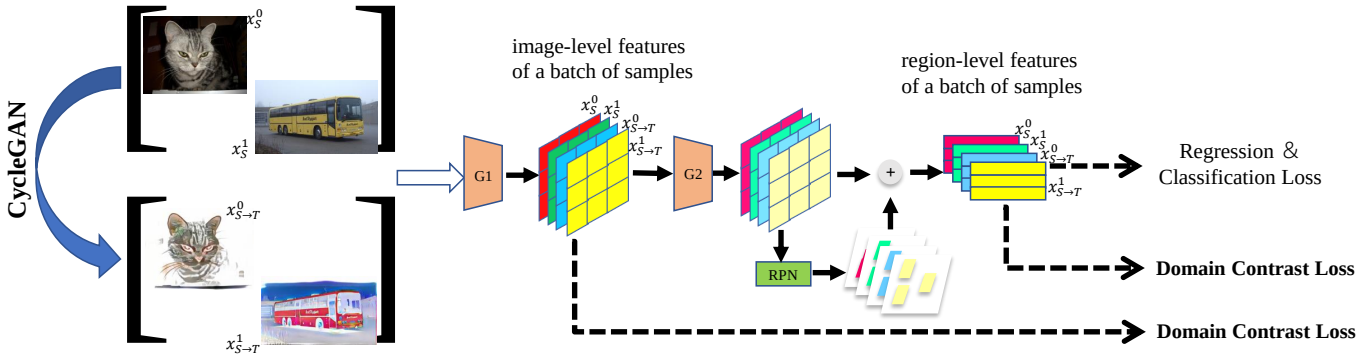


Fig. 2: Illustration of the domain adaptive objector based upon CycleGAN and the proposed DC loss. CycleGAN is adopted to translate an image from the $\mathcal{S}/\mathcal{T}$ to the $\mathcal{T}/\mathcal{S}$ domain. $G1$ and $G2$ refer to convolutional layers. The detection network is first trained within $\mathcal{S}$ by minimizing the training loss $\mathcal{L}_D(\theta)$. The network then is transferred across domains by progressively minimizing the DC loss including $\mathcal{L}^I_{C,\mathcal{S}}(\theta)$, $\mathcal{L}^R_{C,\mathcal{S}}(\theta)$, and $\mathcal{L}^I_{C,\mathcal{T}}(\theta)$ and the pseudo ground-truth loss $\mathcal{L}^R_{\mathcal{T}}(\theta)$.

We therefore have the following inequality[1]

$$-f'(x_t^i)\log(\frac{h^*(x_t^i)}{1-h^*(x_t^i)}) \leq -f'(x_s^i)\log(\frac{h^*(x_s^i)}{1-h^*(x_s^i)}). \quad (6)$$

Substituting Eq. 4 to Eq. 6, we have

$$|\mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{T}}) - \mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{S}})| \leq |\mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}}) - \mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{S}})|. \quad (7)$$

According to Eq. 1, the error bound of hypothesis h^* in the target domain is concluded as

$$\begin{aligned}\mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{T}}) &\leq \mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{S}}) + |\mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}}) - \mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{S}})| \\ &\leq \mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{S}}) + \mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}}).\end{aligned} \quad (8)$$

Domain Contrast (DC) Loss. To quantify $\mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}})$, h^* is supposed to be a nearest neighbor classifier. The probability that a source domain sample x_s^i has the same class label with it's neighbors in the target domains is calculated as their similarity $S(x_s^i,x_t^i)$, where $S(u,v) = u^\top v/||u||||v||$ defines the *cosine* similarity between two samples. Combining the probabilities to a softmax function, the nearest neighbor classifier h^* is defined as

$$h^*(x_t^i) = \frac{\sum_j f_{\mathcal{S}}(x_s^j)\exp(S(x_t^i,x_s^j))}{\sum_j \exp(S(x_t^i,x_s^j))}, \quad (9)$$

where x_t^i is the i^{th} sample transferred from the source to the target domain. Accordingly, minimization of the empirical error of a source model in the target domain, $\mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{S}})$, can be implemented by minimizing

$$\begin{aligned}\mathcal{R}_{\mathcal{T}}(h^*,f_{\mathcal{S}}) &= \frac{1}{N}\sum_i -\log\left(\frac{\sum_j I(x_t^i,x_s^j)\exp(S(x_t^i,x_s^j))}{\sum_j \exp(S(x_t^i,x_s^j))}\right) \\ &\leq \frac{1}{N}\sum_i -\log\left(\frac{\exp(S(x_t^i,x_s^i))}{\sum_j \exp(S(x_t^i,x_s^j))}\right),\end{aligned} \quad (10)$$

where $I(x_1,x_2) = 1-|f_{\mathcal{S}}(x_1)-f_{\mathcal{S}}(x_2)|$. N denotes the number of samples[2]. Correspondingly, we approximately quantify the $\mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}})$ as

$$\mathcal{R}_{\mathcal{S}}(h^*,f_{\mathcal{T}}) \leq \frac{1}{N}\sum_i -\log\left(\frac{\exp(S(x_s^i,x_t^i))}{\sum_j \exp(S(x_s^i,x_t^j))}\right). \quad (11)$$

According to Eqs. 8, 10 and 11, minimizing the error bound defined by Eq. 1 can be fulfilled by optimizing network parameter to minimize

$$\begin{aligned}\mathcal{L}_C(\mathcal{S},\mathcal{T}) = &-\frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_t^i,x_s^i))}{\sum_j \exp(S(x_t^i,x_s^j))}\right) \\ &-\frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_s^i,x_t^i))}{\sum_j \exp(S(x_s^i,x_t^j))}\right),\end{aligned} \quad (12)$$

which is referred to as the DC loss.

B. Domain Adaptive Detection

To implement domain adaptive detection, a base detector based on the deep network is first trained using the annotated data $\{x_s^n, y_s^n\}_{n=1}^{N^s}$ in the source domain by minimizing the detection loss $\mathcal{L}_D(\theta)$ where θ denotes network parameters. The trained base detector is then transferred to the target domain by minimizing the image- and region-level DC loss $\mathcal{L}^I_{C,\mathcal{S}}(\theta)$ and $\mathcal{L}^R_{C,\mathcal{S}}(\theta)$. In turn, the detector is transferred to the source domain by minimizing the image-level DC loss $\mathcal{L}^I_{C,\mathcal{T}}(\theta)$ and fine-tuning the detector using $\mathcal{L}^R_{\mathcal{T}}(\theta)$, which is the pseudo ground-truth loss.

Base Detector. The Faster R-CNN [1] is employed as the base detector, which consists of three stages: convolutional feature extraction, region proposal generation (RPN) and bounding box regression, as shown in Fig. 2. Each input image is represented as image-level features and RPN generates object region proposals, of which region-level features are extracted by ROI-pooling. With region-level features, the category labels and bounding boxes are predicted by classification and regression subnets. Detection loss $\mathcal{L}_D(\theta)$ is composed of the loss of the RPN and the loss of the subnets.

[1]Proof of Eq. 6 is included in Appendix A.

[2]Proof of Eq. 10 is included in the Appendix B.

Detector Transfer. To transfer the Faster RCNN detector parameterized by θ, we first translate each sample image x_s^i from $\mathcal{S}$ to $\mathcal{T}$ using the CycleGAN method [14]. CycleGAN learns to translate an image from a source domain $\mathcal{S}$ to a target domain $\mathcal{T}$ with unpaired examples. Its goal is to learn a mapping $G : \mathcal{S} \rightarrow \mathcal{T}$ such that the distribution of images from $G(\mathcal{S})$ is indistinguishable from the distribution $\mathcal{T}$ with respect to an adversarial loss. During training, an inverse mapping $F : \mathcal{T} \rightarrow \mathcal{S}$ is also learned. Besides, CycleGAN introduces a cycle consistency loss to enforce $F(G(\mathcal{S})) \approx \mathcal{S}$ and $G(F(\mathcal{T})) \approx \mathcal{T}$. After completing the training of the two mapping functions, we convert source domain images to target domain style and convert the target domain images in source domain style.

Denote the features of a translated sample as $x_{s\rightarrow t}^i(\theta)$. For a batch of samples in $\mathcal{S}$ and their translated counterparts in $\mathcal{T}$, we construct a positive sample pair $(x_s^i(\theta), x_{s\rightarrow t}^i(\theta))$ and $N-1$ negative sample pairs $(x_s^j(\theta), x_{s\rightarrow t}^i(\theta))$. Given positive and negative sample pairs, $\mathcal{S} \rightarrow \mathcal{T}$ transfer is implemented by fine-tuning the network parameter to minimize DC loss. By introducing a temperature parameter τ to Eq. 12, the image-level $\mathcal{S} \rightarrow \mathcal{T}$ DC loss is defined as

$$\mathcal{L}_{C,\mathcal{S}}^I(\theta) = -\frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_{s\rightarrow t}^i(\theta), x_s^i(\theta))/\tau)}{\sum_j \exp(S(x_{s\rightarrow t}^i(\theta), x_s^j(\theta))/\tau)}\right) - \frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_s^i(\theta), x_{s\rightarrow t}^i(\theta))/\tau)}{\sum_j \exp(S(x_s^i(\theta), x_{s\rightarrow t}^j(\theta))/\tau)}\right), \tag{13}$$

where $x^i(\theta)$ denotes the features in the last convolutional layer for a sample image.

In each image, we use ground-truth bounding boxes to crop feature maps to guarantee that features are from the same region. The region-level features are denoted as $r^i(\theta)$. The region-level contrast loss $\mathcal{L}_{C,\mathcal{S}}^R(\theta)$ can be calculated by replacing the image-level features $x^i(\theta)$ with the region-level features $r^i(\theta)$.

In a similar way the image-level $\mathcal{T} \rightarrow \mathcal{S}$ DC loss is defined as

$$\mathcal{L}_{C,\mathcal{T}}^I(\theta) = -\frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_t^i(\theta), x_{t\rightarrow s}^i(\theta)/\tau)}{\sum_j \exp(S(x_t^i(\theta), x_{t\rightarrow s}^j(\theta)/\tau)}\right) - \frac{1}{N}\sum_i \log\left(\frac{\exp(S(x_{t\rightarrow s}^i(\theta), x_t^i(\theta)/\tau)}{\sum_j \exp(S(x_{t\rightarrow s}^i(\theta), x_t^j(\theta)/\tau)}\right), \tag{14}$$

As there is no ground-truth object annotated in the target domain, the region-level transfer can not be directly applied. We thereby first translate all the training image from $\mathcal{T}$ to $\mathcal{S}$ using the CycleGAN method. We then use the detector trained in $\mathcal{S}$ to detect high-scored regions in the translated images as pseudo ground-truth objects. With the pseudo ground-truth objects, region-level $\mathcal{T} \rightarrow \mathcal{S}$ transfer is carried out by fine-turning the detector to minimize the detection loss $\mathcal{L}_{\mathcal{T}}^R(\theta)$ for pseudo ground-truth objects.

Discussion. The DC loss for detector transfer is derived from the perspective of error bound minimization, while reflecting the similarity and dis-similarity (contrast) between samples across the domains. By combining the CycleGan with DC loss, we implement detectors' transferability while maintaining their discriminability. Since most negative pairs in a batch have different class labels with the positive pair, minimizing DC loss drives learning feature representations which capture information shared by source and target domains but that are discriminative, $i.e.$, different samples/instances in the two domains have small similarities, $i.e.$, large $cosine$ distances, Fig. 3.

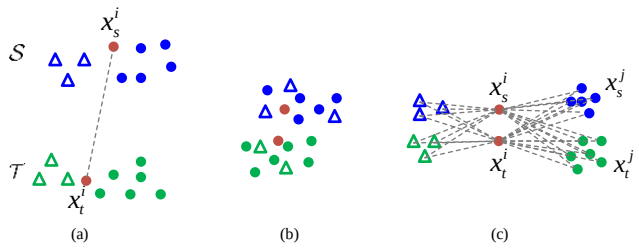


Fig. 3: DC Effect. (a) The reduction of domain divergence (distance between x_s^i and x_t^i) towards aligning feature distributions of the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$. (b) The alignment of $\mathcal{S}$ and $\mathcal{T}$ could reduce the discriminability of models, $i.e.$, samples from different categories are mixed together. (c) Domain Contrast improve the discriminability of models while preserving the alignment of domains.

It is known that the positive-negative class imbalance is an important issue for object detection. Such an issue has been widely explored in supervised detection, but unfortunately ignored by the domain adaptive detection, which could deteriorate the discriminability of transferred detectors. The nominator and denominator of the DC loss naturally incorporate sampling imbalance, which, during transfer, facilities alleviating class imbalance in the target domain. This is an advantage of DC loss compared to the Triplet Loss [45].

IV. Experiment

We analyzed the proposed domain adaptation method by training a detector in the PASCAL VOC dataset [51] and transferring it to the Clipart1K dataset [50] for detection performance evaluation. We also transferred detectors across domains, Pascal VOC → Comic2K, PASCAL → Water-Color2K [50], and SIM 10K[52] → Cityscape[53], where large domain divergence exists.

A. Experimental Setting

Detector and Images. Faster R-CNN with the VGG16[54] or ResNet101 [55] backbone pre-trained on the ImageNet [56] was employed as the base detector. We set the shorter side of the image to 600 pixels following the setting of Faster RCNN [1]. While training the domain adaptive detector, the inputs were a mini-batch of image/frame pairs, including N (batch size) annotated images/frames from the source/target domain and N transferred images. To obtain transferred images, we train CycleGAN [14] with a learning rate of 0.0002 for the first 10 epochs and a linear decaying rate to zero over the next 10 epochs. We followed [14] for other hyper-parameter settings when training the CycleGAN.

TABLE I: Ablation study of detection performance (mAP%) and comparison with the state-of-the-arts when transferring a detector trained within PASCAL VOC to Clipart1K.

Method	aero	bike	bird	boat	bott.	bus	car	cat	chair	cow	table	dog	hrs	mbi.	pers.	plant	she.	sofa	train	tv	mAP
WST-BSR[46]	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
SWDA[47]	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
ICR-CCR [48]	28.7	55.3	**31.8**	26.0	40.1	**63.6**	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN[11]	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3
DM[22]	25.8	63.2	24.5	**42.4**	47.9	43.1	37.5	9.1	**47.0**	46.7	26.8	**24.9**	**48.1**	78.7	63.0	45.0	21.3	36.1	52.3	**53.4**	41.8
ATF[49]	41.9	**67.0**	27.4	36.4	41.0	48.5	42.0	13.1	39.2	**75.1**	**33.4**	7.9	41.2	56.2	61.4	50.6	**42.0**	25.0	53.1	39.1	42.1
Baseline[1]	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
$DC^{I}_{\mathcal{S}\to\mathcal{T}}$	44.0	49.4	34.9	34.0	40.1	52.4	42.2	11.5	38.4	37.1	30.4	15.6	34.0	84.6	58.5	50.2	14.4	24.5	35.6	42.3	38.7
$DC^{R}_{\mathcal{S}\to\mathcal{T}}$	29.4	53.2	27.4	26.4	45.2	51.5	41.0	5.2	35.8	36.5	22.3	9.8	31.7	79.1	51.6	42.3	12.5	25.3	43.6	41.2	35.5
$DC^{I,R}_{\mathcal{S}\to\mathcal{T}}$	40.3	58.7	33.0	31.9	**49.3**	51.9	47.2	6.5	36.8	38.3	32.1	16.7	32.2	85.3	57.9	48.0	15.0	25.9	46.3	44.2	39.9
$DC^{I}_{\mathcal{T}\to\mathcal{S}}$	36.0	53.1	29.9	24.6	40.1	51.0	33.9	7.5	39.2	23.8	23.2	11.5	31.4	59.7	41.9	49.2	10.9	**30.4**	47.9	41.1	34.3
$DC^{I,R}_{\mathcal{S}\to\mathcal{T}}$-$DC^{I}_{\mathcal{T}\to\mathcal{S}}$	45.2	55.9	33.8	32.8	49.2	52.2	48.2	9.4	37.6	38.7	31.8	16.6	34.9	**87.3**	60.3	50.2	15.8	27.4	45.5	47.9	41.0
$DC^{I,R}_{\mathcal{S}\to\mathcal{T}}$- $DC^{I,R}_{\mathcal{T}\to\mathcal{S}}$	**47.1**	53.2	**38.8**	37.0	46.6	45.8	**52.6**	**14.5**	39.1	48.4	31.7	23.7	34.9	87.0	**67.8**	**54.0**	22.8	23.8	44.9	51.0	**43.2**
Oracle	30.1	51.4	47.2	42.5	30.7	55.7	59.4	25.1	47.4	52.5	37.8	43.3	42.6	61.6	73.3	41.9	44.3	25.5	59.0	51.3	46.1

TABLE II: Performance comparison when transferring detectors from Pascal VOC to Comic2k.

Method	Base Detector	bike	bird	car	cat	dog	per.	mAP
DT[50]	SSD+VGG16	43.6	13.6	30.2	16.0	26.9	48.3	29.8
WST-BSR[46]	SSD+VGG16	50.6	13.6	31.0	7.5	16.4	41.4	26.9
DM[22]	Faster RCNN+VGG16	-	-	-	-	-	-	34.5
Baseline[1]	Faster RCNN+VGG16	38.5	10.5	14.7	15.1	15.2	29.8	20.6
	Faster RCNN+ResNet101	30.7	13.8	24.2	13.8	14.8	32.7	21.7
DC (ours)	Faster RCNN+VGG16	**52.7**	17.4	**43.4**	23.3	25.9	58.7	36.9
	Faster RCNN+ResNet101	51.9	**23.9**	36.7	**27.1**	**31.5**	**61.0**	**38.7**
Oracle	Faster RCNN+VGG16	40.1	23.1	32.8	36.7	36.6	68.4	39.6
	Faster RCNN+ResNet101	38.3	30.8	34.9	51.8	47.5	72.8	46.0

Training Details. The detection network (base detector) was trained with a learning rate of 0.001 in the first 5 epochs and decreased to 0.0001 in the following 2 epochs. The iteration number of each epoch is calculated by the sample number divided by the batch size. The mean Average Precision (mAP) for all object categories was used as the evaluation metric. The DC loss is plug-and-play, which means it operates simply by fine-tuning the base detector trained in the source domain. For DC loss, when training with source images and transferred source images, the detector is fine-tuned for 3 epochs with a learning rate of 0.00001. When training using target images and transferred images, the detector is fine-tuned for 3 epochs with a learning rate of 0.00005. The batch size of these two steps is 8. For target images with pseudo labels, the detector is fine-tuned for 3 epochs with a learning rate of 0.00001. In this way, there is no regularization factor required to balance the importance of each loss terms defined in Sec. 3.2, which simplifies the parameter settings.

Baseline and Compared Methods. Faster R-CNN [1] trained only with source domain data in a supervised way is set as our baseline method. All the compared methods, e.g., WST-BSR [46], SWDA [4], ICR-CCR[48], HTCN[11], DM[22], ATF [49], CST [57], leverage adversarial learning to align feature representation of the source domain and the target domain. Besides, WST-BSR [46] and DT [50] generate pseudo labels for target domain images to boost performance. Our model handles the domain discrepancy by using the proposed Domain Contrast Loss. It also leverages pseudo labels in the target domain.

B. Model Effect

Visualization. In Fig. 1, we compared the distributions of target samples before and after domain adaptation. Before adaptation, the target domain samples tends to be concentrated together and are difficult to discriminate. After domain adaptation using DC, the distribution of the target domain samples were well aligned with that of the source domain samples. At the same time, target samples can be well discriminated, which demonstrated the effect of the proposed DC method, demonstrating that it improves detectors' discriminability while maintaining the transferability. The detection examples in Fig.4 also demonstrate the effectiveness of our DC Loss.

Ablation Study. In Table I, the effect of domain contrast was validated by performing $\mathcal{S} \to \mathcal{T}$ and $\mathcal{T} \to \mathcal{S}$ transfer step-by-step. Using solely the image-level $\mathcal{S} \to \mathcal{T}$ transfer ($DC^{I}_{\mathcal{S}\to\mathcal{T}}$) the mAP is improved by 10.9% (38.7% vs. 27.8%) which validated the effectiveness of our approach at reducing cross-domain divergence while improving detector discriminability. The region-level $\mathcal{S} \to \mathcal{T}$ transfer ($DC^{R}_{\mathcal{S}\to\mathcal{T}}$) improved the mAP by 7.7%. Combining the image-level and region-level $\mathcal{S} \to \mathcal{T}$ transfer ($DC^{I,R}_{\mathcal{S}\to\mathcal{T}}$) improved the mAP by 12.1% (39.9% vs. 27.8%).

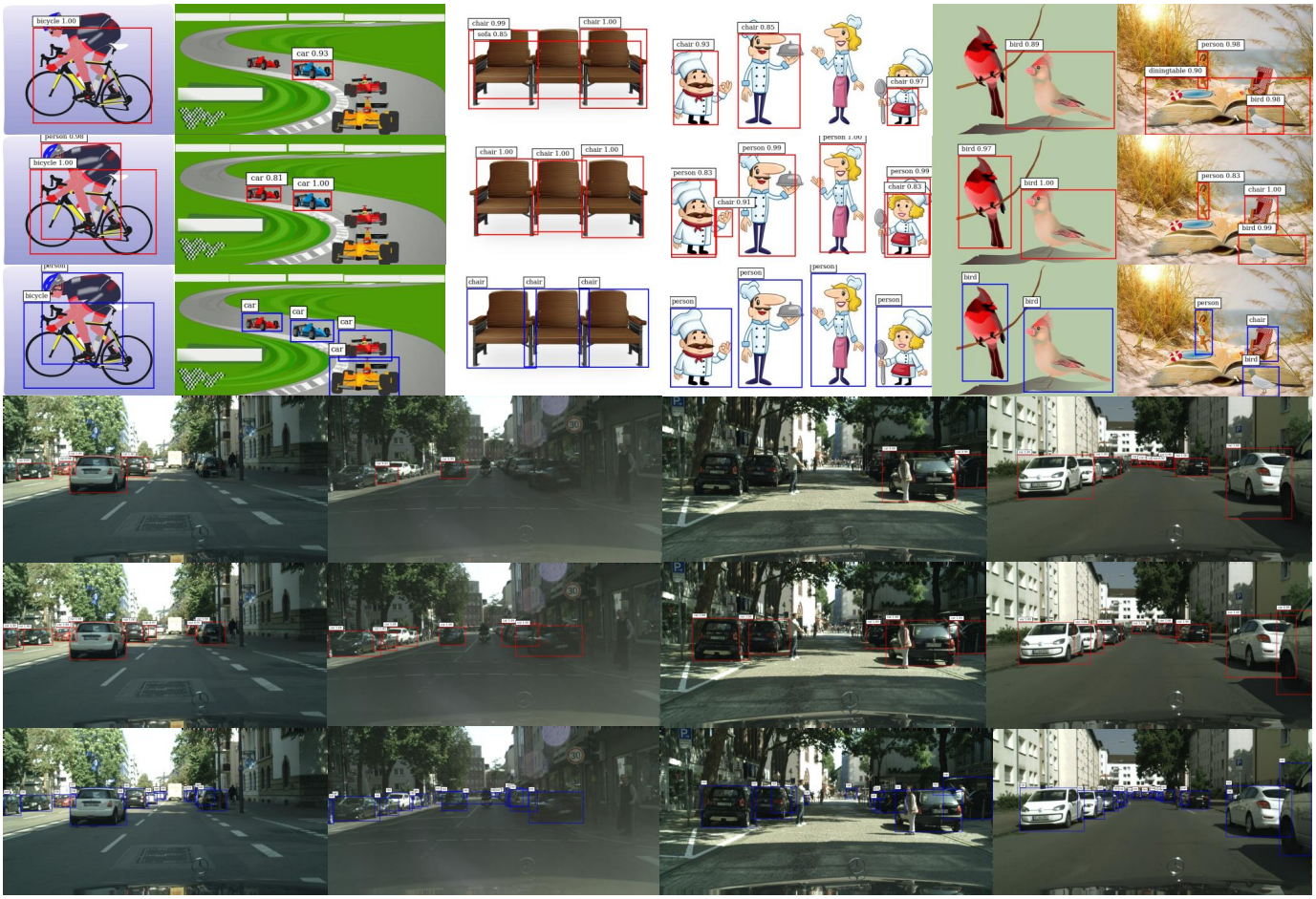


Fig. 4: Detection examples from the Clipart1K and Cityscape datasets. The first three rows are for Clipart1k and the left are for Cityscape. The results of the baseline method are in the first row and our detection results are in the second row. The Ground Truth are in the third row. (Best viewed in color)

After performing image-level $\mathcal{T} \rightarrow \mathcal{S}$ transfer ($\text{DC}^{I,R}_{\mathcal{S}\rightarrow\mathcal{T}}$-$\text{DC}^{I}_{\mathcal{T}\rightarrow\mathcal{S}}$), the mAP is further improved by 1.1% (41.0% vs. 39.9%). Using the pseudo objects in the target domain to fine tune the detector with DC loss was also effective, improving the mAP by 2.2% (43.2% vs. 41.0%). To reduce false positives, the threshold for pseudo object detection was set to be 0.95. Without bells and whistles, our method outperformed the state-of-the-art by 1.1% (43.2% vs. 42.1%), which was a significant margin considering the large domain divergence. We reported the "Oracle" result by training a Faster RCNN detector using the images within target domain but with the ground truth annotations, which is a reference for the performance upper-bound.

Parameter Setting. In Fig. 5, parameters τ and batch size were analyzed for DC loss. With $\tau = 0.5$ the best performance was achieved. The optimization of τ improved the mAP by 1.75%, showing the importance of the temperature parameter. In Fig. 6, the performance was not very sensitive to the batch size, *e.g.*, using a large batch size 8 slightly improved the mAP. This is different from unsupervised contrastive learning which often relies on large batch sizes [39], [58].

Style Transfer Methods. For a fair comparison, we follow DT [50] and use CycleGAN as the default style transfer

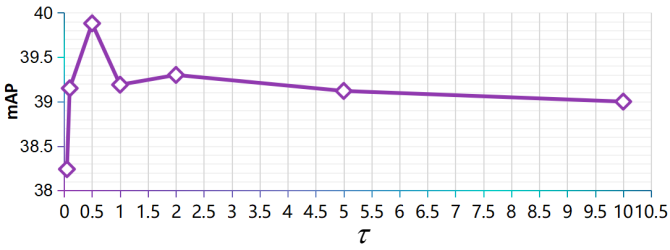


Fig. 5: Ablation of parameter τ at proper learning rates.

method. To further explore the style transfer methods, we test two more style methods. One is AdaIN [63], which uses instance normalization to conduct style transfer, without using adversarial training.The other is StyleNAS [62], which is built upon neural architecture search. As shown in Table VI, both AdaIn and StyleNAS obtain higher detection performance than CyclGAN. This show that our DC method is compatible with different style transfer methods and preserve the advantages of them.

Comparison with Triplet Loss and MMD. In Fig. 7, we demonstrate the advantage of DC loss over the Triplet

TABLE III: Performance comparison when transferring detectors from Pascal VOC to WaterColor2K.

Method	Base Detector	bike	bird	car	cat	dog	per.	mAP
DT[50]	SSD+VGG16	**82.8**	47.0	40.2	34.6	35.3	62.5	50.4
WST-BSR[46]	SSD+VGG16	75.6	45.8	**49.3**	34.1	30.1	64.1	49.9
DM[22]	Faster RCNN+VGG16	-	-	-	-	-	-	52.0
SWDA[47]	Faster RCNN+ResNet101	82.3	55.9	46.5	32.7	**35.5**	66.7	53.3
ATF[49]	Faster RCNN+ResNet101	78.8	**59.9**	47.9	41.0	34.8	66.9	**54.9**
Baseline[1]	Faster RCNN+ResNet101	77.4	46.5	39.7	32.4	24.3	57.5	46.3
DC (ours)	Faster RCNN+ResNet101	76.7	53.2	45.3	**41.6**	**35.5**	**70.0**	53.7
Oracle	Faster RCNN+ResNet101	70.9	52.9	44.4	41.7	51.4	74.5	56.0

TABLE IV: Performance comparison when transferring detectors from SIM 10K to Cityscape.

Method	Base Detector	AP on car
DA Faster[3]	Faster RCNN+VGG16	39.0
SWDA[47]	Faster RCNN+VGG16	40.1
MAF[59]	Faster RCNN+VGG16	41.1
HTCN[11]	Faster RCNN+VGG16	42.5
SCDA[60]	Faster RCNN+VGG16	43.0
CST[57]	Faster RCNN+VGG16	**44.5**
Baseline[1]	Faster RCNN+VGG16	34.2
DC (ours)	Faster RCNN+VGG16	41.6
Oracle	Faster RCNN+VGG16	53.2

TABLE V: Accuracy(%) on Office-Home for unsupervised domain adaptive image classification.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
CDAN[61]-reported	49	69.3	74.5	54.4	66	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN[61]-reproduced	50.7	68.7	74.9	53.8	**69.9**	68.9	56	49.6	75.3	71.5	55.3	80.7	64.6
CDAN[61]+DC$^{I}_{\mathcal{S}\to\mathcal{T}}$	**53**	**71.1**	**75.4**	**58**	**69.9**	**69.3**	**57.8**	**51.8**	**76.7**	**72.2**	**57.7**	**82.3**	**66.3**

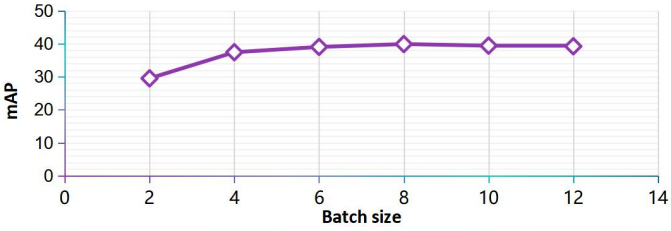


Fig. 6: Ablation of batch size.

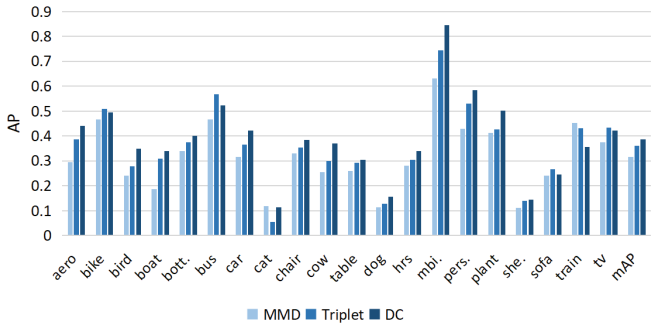


Fig. 7: Comparison of APs and mAP of Triplet Loss, MMD, and DC loss when performing image-level transfer from Pascal VOC to Clipart1K.

TABLE VI: Ablation for style transfer method on image-level $\mathcal{S}\to\mathcal{T}$ transfer in PASCAL VOC to Clipart1K task.

Style Transfer Method	mAP(%)
Baseline[1]	27.8
CycleGAN[14]	38.7
StyleNAS[62]	39.1
AdaIN[63]	40.1

loss [45], which is designed to minimize the intra-class distance of each positive sample pair and maximizes the inter-class distances of each positive-negative sample pair as $\sum_j \max(||x^i_s - x^i_{s\to t}||^2_2 - ||x^i_s - x^j_{s\to t}||^2_2 + 0.5, 0)$. It can be seen that Triplet Loss reported lower mAP as it pursued maximizing the domain similarity and minimizing the similarity of a single pair of sample in a mini-batch but unfortunately ignored the class imbalance issue in object detection. This caused more false detection results from the background areas. Such an imbalance issue was also ignored by the MMD method [13].

C. Performance and Comparison

In Table II, we evaluated the proposed method and compared it with state-of-the-art methods when transferring detectors from Pascal VOC to Comic2k [50]. The Comic2k dataset includes 2,000 comic images, 1,000 for training and the other 1,000 for test. It has 6 object classes which also exist in Pascal

VOC. As the images in Comic2k are unrealistic images, the domain divergence between the source (VOC) and the target domain (Comic2k) was predictably large. In this scenario, the proposed DC method outperformed the state-of-the-art method WST-BSR [46] by 10.0% (36.9% vs. 26.9%) and DM [22] by 2.4% (36.9% vs. 34.5%). Using the ResNet-101 backbone further boosted the performance to 38.7%.

In Table III and Table IV, our method is comparable with, if not outperforms, the state-of-the-art domain adaptive detectors. Compared with the objects in other datasets, the objects in SIM 10k and CityScape typically occupy small areas compared to the whole image. The background interference is thereby more significant when performing detector training. To transfer the detectors in such scenarios it requires more sophisticated region-level domain adaptation methods, which is the future research direction of DC.

D. Generalization Performance on UDA Classification.

To further verify the effectiveness of the DC Loss, we apply it for domain adaptive image classification. We conduct experiments with the classical benchmark CDAN [61] on the challenging Office-Home [64] Dataset. As shown in Table V, with image-level DC Loss, we steadily improve the performance on multiple sub tasks, yielding 2.5% gain on the reported results and 1.7% gain compared to the reproduced results.

V. Conclusion and Future Research

We have presented Domain Contrast (DC), a simple yet effective approach to train domain adaptive detectors. This DC method is theoretically plausible because it was deduced from the perspective of error bound minimization about transfer learning. It is also conceptually simple and can simultaneously guarantee the transferability of detectors while preserving the discriminability of transferred detectors by minimizing DC loss. DC significantly boosted the performance of domain adaptive detectors and improved the state-of-the-art on image and video datasets of large domain divergence. DC not only provides a fresh insight to the transfer learning problem but also a practical technique to handle the large domain divergence issue in object detection and recognition scenarios. As a future research direction, DC Loss can be leveraged to explore the temporal relevance across video frames to boost performance by constructing more positive pairs across adjacent frames.

Acknowledgment

This work was supported by Natural Science Foundation of China (NSFC) under Grant 61836012, 61771447 and 62006216, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000.

Appendix A
Proof of Eq.6 in the paper

$$-f'(x_t^i)\log(\frac{h^*(x_t^i)}{1-h^*(x_t^i)}) \le -f'(x_s^i)\log(\frac{h^*(x_s^i)}{1-h^*(x_s^i)}).$$

Proof
Since $f_{\mathcal{T}}(x_s^i) = f_{\mathcal{T}}(x_t^i)$ and $f_{\mathcal{S}}(x_s^i) = f_{\mathcal{S}}(x_t^i)$,

$$f'(x_t^i) = f_{\mathcal{T}}(x_t^i) - f_{\mathcal{S}}(x_t^i) = f_{\mathcal{T}}(x_s^i) - f_{\mathcal{S}}(x_s^i) = f'(x_s^i).$$

1) $f_{\mathcal{S}}(x) = f_{\mathcal{T}}(x)$, $f'(x) = 0$
Obviously, Eq.6 holds.

2) $f_{\mathcal{S}}(x) = 0$, $f_{\mathcal{T}}(x) = 1$, $f'(x) = 1$
Since h^* is the optimal hypothesis in the source domain,

$$0 \le h^*(x_s^i) \le 1/2,\ log(\tfrac{h^*(x_s^i)}{1-h^*(x_s^i)}) < 0.$$

Since $|\log(\frac{h^*(x_t^i)}{1-h^*(x_t^i)})| \le |\log(\frac{h^*(x_s^i)}{1-h^*(x_s^i)})|$ (Eq. 5) holds, Eq. 6 holds.

3) $f_{\mathcal{S}}(x) = 1$, $f_{\mathcal{T}}(x) = 0$, $f'(x) = -1$
Since h^* is the optimal hypothesis in the source domain,

$$1/2 \le h^*(x_s^i) \le 1,\ log(\tfrac{h^*(x_s^i)}{1-h^*(x_s^i)}) > 0.$$

Since $|\log(\frac{h^*(x_t^i)}{1-h^*(x_t^i)})| \le |\log(\frac{h^*(x_s^i)}{1-h^*(x_s^i)})|$ (Eq. 5) holds, Eq. 6 holds.

Appendix B
Proof of Eq.10 in the paper

$$\begin{aligned}\mathcal{R}_{\mathcal{T}}(h^*, f_{\mathcal{S}}) &= \frac{1}{N}\sum_i -\log\left(\frac{\sum_j I(x_t^i, x_s^j)\exp(S(x_t^i, x_s^j))}{\sum_j \exp(S(x_t^i, x_s^j))}\right) \\ &\le \frac{1}{N}\sum_i -\log\left(\frac{\exp(S(x_t^i, x_s^i))}{\sum_j \exp(S(x_t^i, x_s^j))}\right).\end{aligned}$$

Proof
It's equal to prove that

$$\begin{aligned}&\frac{1}{N}\sum_i -\log(\sum_j I(x_t^i, x_s^j)\exp(S(x_t^i, x_s^j))) \\ \le &\frac{1}{N}\sum_i -\log(\exp(S(x_t^i, x_s^i))).\end{aligned}$$

Following the assumption of co-variate shift, when $i = j$, $I(x_t^i, x_s^j) = 1 - |f_{\mathcal{S}}(x_t^i) - f_{\mathcal{S}}(x_s^j)| = 1$. When $i \ne j$, $0 \le I(x_t^i, x_s^j) \le 1$. Besides, $\exp(S(x_t^i, x_s^j)) > 0$. Therefore,

$$\frac{1}{N}\sum_i\sum_j I(x_t^i, x_s^j)\exp(S(x_t^i, x_s^j)) \ge \frac{1}{N}\sum_i \exp(S(x_t^i, x_s^i)).$$

Since function $\log(\cdot)$ increases monotonically,

$$\begin{aligned}&\frac{1}{N}\sum_i -\log(\sum_j I(x_t^i, x_s^j)\exp(S(x_t^i, x_s^j))) \\ \le &\frac{1}{N}\sum_i -\log(\exp(S(x_t^i, x_s^i))).\end{aligned}$$

REFERENCES

[1] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99. 1, 4, 5, 6, 8

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *ECCV*, vol. 9905, 2016, pp. 21–37. 1

[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *IEEE CVPR*, 2018, pp. 3339–3348. 1, 2, 3, 8

[4] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *IEEE CVPR*, 2019, pp. 6956–6965. 1, 2, 3, 6

[5] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *IEEE CVPR*, 2019, pp. 687–696. 1, 2, 3

[6] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *IEEE CVPR*, 2019, pp. 7289–7298. 1

[7] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *IEEE CVPR*, 2019. 1

[8] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE CVPR*, 2018. 1

[9] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *IEEE CVPR*, 2018. 1

[10] S. Lin, H. Li, C. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," in *BMVC*, 2018. 1

[11] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *IEEE CVPR*, 2020. 1, 3, 6, 8

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012. 1, 2

[13] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *IEEE ICCV*. IEEE, 2019, pp. 1426–1435. 1, 8

[14] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE ICCV*, 2017, pp. 2242–2251. 1, 5, 8

[15] W. Li, Z. Xu, D. Xu, D. Dai, and L. V. Gool, "Domain generalization and adaptation using low rank exemplar svms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1114–1127, 2018. 2

[16] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *IEEE ICCV*, 2017, pp. 5715–5725. 2

[17] P. P. Busto and J. Gall, "Open set domain adaptation," in *IEEE ICCV*, 2017, pp. 754–763. 2

[18] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, vol. 37, 2015, pp. 1180–1189. 2

[19] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, F. R. Bach and D. M. Blei, Eds., vol. 37, 2015, pp. 97–105. 2

[20] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE CVPR*, 2018, pp. 994–1003. 2

[21] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *IEEE CVPR*, 2019, pp. 598–607. 2

[22] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *IEEE CVPR*, 2019, pp. 12 456–12 465. 2, 3, 6, 8, 9

[23] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *NeurIPS*, 2016, pp. 136–144. 2

[24] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang, "Progressive feature alignment for unsupervised domain adaptation," in *IEEE CVPR*, 2019, pp. 627–636. 2

[25] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *IEEE CVPR*, 2018. 2

[26] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6321–6330. 2

[27] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *IEEE CVPR*, 2017, pp. 95–104. 2

[28] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *ICML*, 2018, pp. 1994–2003. 2

[29] J. Choi, G. Sharma, S. Schulter, and J.-B. Huang, "Shuffle and attend: Video domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 678–695. 2

[30] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 815–11 822. 2

[31] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 122–132. 2

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742. 2

[33] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742. 2

[34] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304. 2

[35] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 2

[36] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018. 2

[37] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019. 2

[38] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 535–15 545. 2

[39] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *CoRR*, vol. abs/1906.05849, 2019. 3, 7

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 3

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020. 3

[42] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902. 3

[43] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *IEEE CVPR*, 2019, pp. 11 457–11 466. 3

[44] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1-2, pp. 151–175, 2010. 3

[45] P. Laiz, J. Vitrià, and S. Seguí, "Using the triplet loss for domain adaptation in WCE," in *IEEE ICCV Workshop*, 2019, pp. 399–405. 5, 8

[46] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial bakground regularization for unsupervised domain adaptive one-stage object detection," in *IEEE ICCV*, 2019, pp. 6091–6100. 6, 8, 9

[47] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *IEEE CVPR*, 2019, pp. 6956–6965. 6, 8

[48] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *IEEE CVPR*, 2020. 6

[49] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," *arXiv preprint arXiv:2007.01571*, 2020. 6, 8

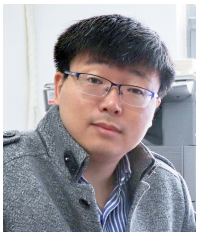
[50] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *IEEE CVPR*, 2018, pp. 5001–5009. 5, 6, 7, 8
[51] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010. 5
[52] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" *arXiv preprint arXiv:1610.01983*, 2016. 5
[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE CVPR*, 2016, pp. 3213–3223. 5
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778. 5
[56] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255. 5
[57] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 86–102. 6, 8
[58] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. 7
[59] Z. He and L. Zhang, "Multi-adversarial faster-rcnn for unrestricted object detection," in *IEEE ICCV*, 2019, pp. 6668–6677. 8
[60] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *IEEE CVPR*, 2019, pp. 687–696. 8
[61] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *NeurIPS*, 2018. 8, 9
[62] J. An, H. Xiong, J. Huan, and J. Luo, "Ultrafast photorealistic style transfer via neural architecture search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 443–10 450. 7, 8
[63] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. 7, 8
[64] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027. 9

Feng Liu received the B.S. degree from Harbin Institute of Technology, China, in 2020. He has been a master student in the School of Microelectronics, University of Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and machine learning, specifically for visual object detection.

Xiaosong Zhang received the B.S. degree from Harbin Institute of Technology (HIT), Weihai, Shandong, China, in 2018. Since 2018, he has been a Ph.D student in the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and machine learning, specifically for visual object detection and representation learning.

Fang Wan received the B.S. degree from Wuhan University, Wuhan, China, in 2013. Since 2013, he has been a Ph.D student in the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences. His research interests include computer vision and machine learning. He has published more than 10 papers in refereed conferences and journals including IEEE CVPR, ICCV, NeurIPS, and PAMI, and received President Award of Chinese Academy of Sciences.

Xiangyang Ji (M'10) received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He joined Tsinghua University, Beijing, in 2008, where he is currently a Professor with the Department of Automation. He has authored over 100 referred conference and journal papers. His current research interests include signal processing, image/video compressing, and intelligent imaging.



Qixiang Ye (M'10-SM'15) received the B.S. and M.S. degrees from Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2006. He has been a professor with the University of Chinese Academy of Sciences (UCAS) since 2015, and was a visiting assistant professor with the University of Maryland, College Park until 2013. His research interests include visual object detection and machine learning. He has published more than 100 papers in refereed conferences and journals including IEEE CVPR, ICCV, ECCV, NeurIPS, and PAMI. He is a senior member of IEEE.