Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Object Detection

Taekyung Kim Minki Jeong Seunghyeon Kim Seokeon Choi Changick Kim
Korea Advanced Institute of Science and Technology, Daejeon, Korea
{tkkim93, rhm033, seunghyeonkim, seokeon, changick}@kaist.ac.kr

Abstract

We introduce a novel unsupervised domain adaptation approach for object detection. We aim to alleviate the imperfect translation problem of pixel-level adaptations, and the source-biased discriminativity problem of feature-level adaptations simultaneously. Our approach is composed of two stages, i.e.*, Domain Diversification (DD) and Multi-domain-invariant Representation Learning (MRL). At the DD stage, we diversify the distribution of the labeled data by generating various distinctive shifted domains from the source domain. At the MRL stage, we apply adversarial learning with a multi-domain discriminator to encourage feature to be indistinguishable among the domains. DD addresses the source-biased discriminativity, while MRL mitigates the imperfect image translation. We construct a structured domain adaptation framework for our learning paradigm and introduce a practical way of DD for implementation. Our method outperforms the state-of-the-art methods by a large margin of 3% $\sim$ 12% in terms of mean average precision (mAP) on various datasets.*

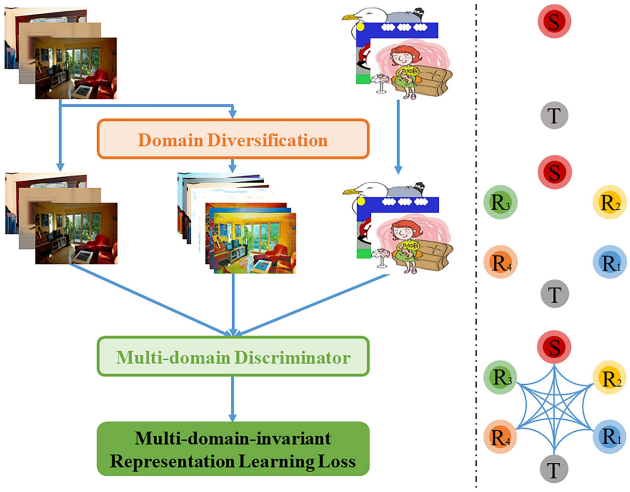


Figure 1. Overview of our learning paradigm. We illustrate a conceptual diagram of the distributions of the domains on the right side. S and T represent for the source and the target domain, respectively, and each R_i represents the ith diversified domain.

1. Introduction

Object detection is a fundamental problem in computer vision as well as machine learning. With the recent advances of the convolutional neural networks (CNNs), CNN-based methods [13, 12, 35, 30, 34, 26, 8, 46, 29] have achieved significant progress in object detection based on fine benchmarks [10, 27, 25]. Despite the promising results, all of these object detectors suffer from the degenerative problem when applied beyond these benchmarks. Building datasets for a specific application can temporarily resolve this problem, nevertheless, the time and monetary costs incurred when manually annotating such datasets are not negligible [40, 33]. Moreover, since the intrinsic causes of the degenerative problem have been avoided instead of resolved, another generalization issue arises when extending the same application to different environments. To address this issue, an unsupervised domain adaptation method for object detection [3] was recently proposed.

Unsupervised domain adaptation has been studied to address the degeneration issue between related domains, which is closely related to the aforementioned degenerative problem. With the rise of the deep neural networks, recent unsupervised deep domain adaptation methods [31, 11, 42, 2, 36, 1, 17] are mainly based on feature-level adaptation and pixel-level adaptation. Feature-level adaptation methods [31, 11, 42, 2] align the distributions of the source and the target domain toward a cross-domain feature space. These approaches expect the model supervised by the labeled source domain to infer on the target domain effectively. However, the supervision of the inference layer mainly relies on the source domain only in the feature-level adaptation methods. Thus, the feature extractor of the model is enforced to manufacture the features in a way discriminative for the source domain data, which is not suitable

for the target domain. Moreover, since the object detection data is interwoven with the instances of interest and the relatively unimportant background, it is further hard for the source-biased feature extractor to extract discriminative features for the target domain instances. Thus, object detectors adapted at the feature-level are at risk of the source-biased discriminativity and it can leads to false recognition on the target domain. On the other hand, pixel-level adaptation methods [36, 1, 17] focus on visual appearance translation toward the opposite domain. The model can then take advantage of the information from the translated source images [17, 1] or infer pseudo label of the translated target images [22]. Most existing pixel-level adaptation methods [36, 1, 17] are based on the assumption that the image translator can perfectly convert one domain to the opposite domain such that the translated images can be regarded as those from the opposite domain. However, these methods reveal imperfect translation in many adaptation cases since the performance of the translator heavily depends on the appearance gap between the source and the target domain, as shown in Fig. 2. Regarding these incompletely translated source images as from the target domain can cause new domain discrepancy issue.

To tackle the aforementioned limitations, we introduce a novel domain adaptation paradigm for object detection. Our learning paradigm consists of Domain Diversification (DD) and Multi-domain-invariant Representation Learning (MRL), as shown in Fig. 1. Unlike most existing domain adaptation methods, DD intentionally causes several distinctive shifted domains from the source domain to enrich the distribution of the labeled data. On the other hand, MRL boosts the domain invariance of the features by unifying the scattered domains. Using the aforementioned approaches, we propose a universal domain adaptation framework for object detection. Our framework trains domain-invariant object detection layers with diversified annotated data while simultaneously encouraging dispersed domains toward a common feature space. To demonstrate the effectiveness of our method, we conduct extensive experiments on Real-world Datasets [10], Artistic Media Datasets [22], and Urban Scene Datasets [7, 37] based on Faster R-CNN. Our framework achieves state-of-the-art performance on various datasets.

In summary, we have three contributions in our paper:

- We propose a novel learning paradigm for unsupervised domain adaptation. Our learning approach addresses the source-biased discriminativity issue and the imperfect translation issue.
- We structurize our learning paradigm by integrating DD and MRL in the form of a framework.
- We conduct extensive experiments to validate the effectiveness of our method on various datasets. Our method outperforms the state-of-the-art methods with a large margin by $3\% \sim 12\%$ mAP.

(a) Source domain (b) Target domain (c) Translated domain

Figure 2. Examples of the imperfect image translation. The first and second rows visualize examples of the translated image from the real-world to artistic media and between urban scenes, respectively.

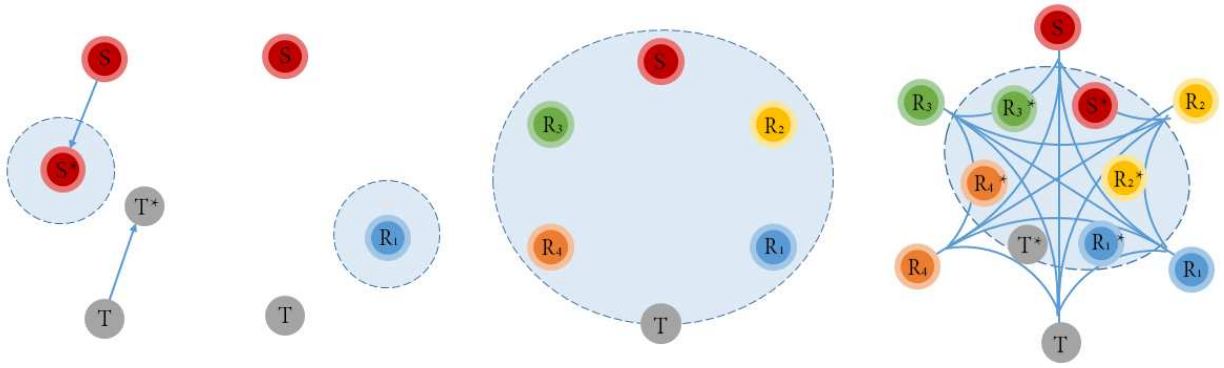
2. Related work

2.1. CNN-based Object Detection

Traditional methods [44, 9] use a sliding window framework with handcrafted features and shallow inference models. With rise of the convolutional neural networks, R-CNN [13] obtains a promising result with a selective search algorithm and classification through the CNN features. Fast R-CNN [12] reduces the bottleneck of R-CNN by sharing features among regions in the same image. Faster R-CNN [35] adopts a fully convolutional network called a Region Proposal Network (RPN) to mitigate another bottleneck caused by the selective search algorithm. YOLO [34] achieves significant improvement in the inference speed using a single-staged network. SSD [30] uses multi-scale features to enhance the relatively low accuracy of YOLO. RetinaNet [26] further improves the performance of single-staged object detectors using the focal loss to reduces the performance degradation caused by easy negative examples. While these methods push the limit on the large-scale datasets with rich annotations, generalization errors which arises during their application have not been investigated thus far.

2.2. Unsupervised Domain Adaptation

Domain adaptation has been studied intensely in relation to the image classification task [21, 41]. Traditional methods focus on reducing domain discrepancy through instance re-weighting [21, 41, 14] and shallow feature alignment strategies [16, 32]. With the success of deep learning scheme, early deep domain adaptation mainly arises into Maximum Mean Discrepancy (MMD) minimization [31, 42, 2] or feature confusion through adversarial



(a) Feature-level adaptation (b) Pixel-level adaptation (c) Domain Diversification (d) MRL with Domain Diversification

Figure 3. Comparison of distribution transformation by different domain adaptation methods. MRL refers to Multi-domain-invariant Representation Learning. S and T denote the source domain and the target domain, respectively. R_1, R_2, R_3, and R_4 are shifted domains of the source domain. The arrows indicate the feature-level adaptation trends. The domains with asterisks denote the results of feature-level adaptation. The domains with a boundary imply that the object detection network is supervised by these domains.

learning [11]. Recently, as the image-to-image translation has become highlighted with promising results [23, 24, 28, 49] through Generative Adversarial Networks (GANs) [15], pixel-level adaptation methods [36, 20, 1] have been developed to address the domain shift issue by translating source domain images into the target style. As unsupervised domain adaptation attracted considerable interest with its effectiveness, recent works [17, 47, 6, 5, 38, 43, 19, 48] have been attempted to address the generalization issue in the semantic segmentation task.

Despite the recent success of unsupervised domain adaptation in various computer vision tasks, unsupervised domain adaptation for the object detection task has not been explored so far except few pioneers [22, 3]. Inoue et al. [22] adopt a conventional unsupervised pixel-level domain adaptation method as part of a two-staged weakly supervised domain adaptation framework. Chen et al. [3] align distributions of the source and the target domain at the image level and instance level to address various causes of the domain shift separately. While these methods address the problem of degeneracy without considering the limitations of existing domain adaptation approaches, we aim to mitigate these issues through a two-step learning paradigm.

3. Methods

We propose a novel learning paradigm to alleviate the source-biased discriminativity in feature-level adaptation and the imperfect translation in pixel-level adaptation. We start by explaining the two stages of our method, Domain Diversification and Multi-domain-invariant Representation Learning. Then, a universal domain adaptation framework for object detection is introduced. Figure 3 shows conceptual description of feature-level adaptation, pixel-level adaptation, and our method.

(a) Given image (b) Images with appearance shift

Figure 4. Examples of variously shifted images for given images.

3.1. Domain Diversification

Without loss of generality, we assume that there exist numerous possibilities of shifted domains that preserve the corresponding semantic information of the source domain but appear in different ways. For instance, as shown in Fig. 4, we can easily conceive of various visually shifted images from a given image regardless of the existence of a feasible image translator. Along the same line, numerous variations of image translators can achieve considerable domain shift from the given source domain, which we call domain shifters. Domain Diversification (DD) is a method which diversifies the source domain by intentionally generating distinctive domain discrepancy through these domain shifters. The diversified distribution of the labeled data encourages the model to infer among data with large intra-class variance discriminatively. Thus, the model is enforced to extract semantic features that are not biased to a particular domain. This allows the model to extract unbiased semantic features from the target domain, which is more discriminative than the source-biased features. With the better discriminativity of target domain features, we can assimilate the domains with less feature collapse, resulting in more

desirable adaptation.

Among the plenteous possibilities of domain shifters, inspired by the limitation of pixel-level adaptation, we practically realize the possibilities using the imperfections of the image translation. Let us denote a source domain sample as x^s and a target domain sample as x^t with domain distributions p_s and p_t, respectively. In general, image translation methods aim to train a generator G by optimizing the translated image $G(x^s)$ to which appears to be sampled from the target domain. However, since the generator network has high enough capacity for various translations, the adversarial loss alone cannot guarantee the conversion of a given x^s to the desired target image. To redeem this instability, image translation methods add constraints to the objective function L_{im} to reduce the possibility of the undesirable generators:

$$L_{\text{im}}(G, D, M) = L_{\text{GAN}}(G, D) + \alpha L_{\text{con}}(G, M), \quad (1)$$

$$L_{\text{GAN}}(G, D) = \mathbb{E}_{x^t \sim p_{\text{t}}(x^t)}[logD(x^t)] + \mathbb{E}_{x^s \sim p_{\text{s}}(x^s)}[log(1 - D(G(x^s)))], \quad (2)$$

where D is the discriminator for adversarial learning, $L_{con}(G, M)$ is the constraint loss with a possibly existing additional module M and α is a weight that balances the two losses. Here, the additional module implies a supplemental network necessary for a sophisticated constraint.

In this basic setting, we observe that varying the learning trend with alternative constraints causes the generator G to diversify the appearance of the translated images. Based on this observation, we apply several variants of constraints to achieve distinct domain shifters. The objective function for the domain shifter can be written as:

$$L_{\text{DS}}(G, D, M) = L_{\text{GAN}}(G, D) + \beta L_{\text{con}}(G, M), \quad (3)$$

where $L_{con}(G, D, M)$ is the loss for constraints that encourages the domain shifter to be differentiated, M denotes possibly existing additional modules for the constraint loss, and β is a weight that balances the two losses. Practical implementation details for diversifying domain shifters will be introduced in section 4.2.

3.2. Multi-domain-invariant Representation Learning

In conventional pixel-level adaptations, substantial training of the inference layer heavily depends on the translated source images. However, these methods run the risk of imperfect image translation, which can cause another domain shift issue with the target domain. To address this limitation, we design an adversarial learning scheme called Multi-domain-invariant Representation Learning (MRL), which encourages domain-invariant features among the diversely scattered domains through adversarial learning. We assume that we have $(n + 2)$ number of diversified domains with a pairwise domain gap, following the pixel-level adaptation methods. For instance, we regard the translated source domain as separate from the source or the target domain and consider the three domains for conventional pixel-level adaptation methods. In most existing feature-level adaptation methods, the adversarial learning is applied through the binary discriminator. However, these domains have pairwise domain shifts given by the domain adaptation problem or caused by the imperfect image translation. Thus, regarding multiple domains as the same domain during adversarial learning can fatally disturb the model from learning common features. Thus, we use the discriminator with $(n + 2)$ outputs so as to learn to distinguish the domains using the cross entropy loss.

Adversarial learning methods attain domain-invariant features by inducing a feature which confuses the domain discriminator. Thus, in conventional cross-domain adaptation problems, confusion in the discriminator can be achieved by designating each domain to resemble the other. However, in a multi-domain situation, it is not desirable to specify each domain to resemble each specific target domain. To address this issue, inspired by [11], we attach a gradient reverse layer (GRL) at the front-end of the discriminator. Since the GRL forces the generator to manufacture the features of the given images as if they were not sampled from its domain, the features of each domain are encouraged to be domain-invariant. The objective function for MRL can be written as:

$$L_{\text{mrl}}(x^f, D_{x^f}) = -\sum_{i=0}^{n+1} \sum_{u,v} \mathbf{1}_{\{i\}}(D_{x^f}) log(p_i^{(\text{u,v})}(x^f)) \quad (4)$$

where x^f is the feature map given for the discriminator, $1_{\{i\}}$ is the indicator function for a singleton $\{i\}$, $p_i^{(u,v)}$ is the domain probability for the ith domain of the feature vector located at (u, v) of x^f, and D_{x^f} is the ground-truth for the domain label of x^f.

3.3. Structured Domain Adaptation framework for Object Detection

In this section, we structurize our learning paradigm by integrating DD and MRL into a framework. Without loss of generality, we assume that there is n number of domain shifters G_i for $i = 1, ..., n$. Our framework aims to learn domain-invariant representation and adapt the object detector for these representations simultaneously. To achieve the goal, every $(n+2)$ number of domains is utilized for MRL, while the source domain and the shifted domains encourage the localization layers and the classification layers of the object detector. The objective function for the framework can be written as follows:

$$\mathcal{L}(x^s, x^t, y^s) = L_{\text{MRL}}(x^s, x^t) + L_{\text{LOC}}(x^s, y^s) + L_{\text{CLS}}(x^s, y^s), \quad (5)$$

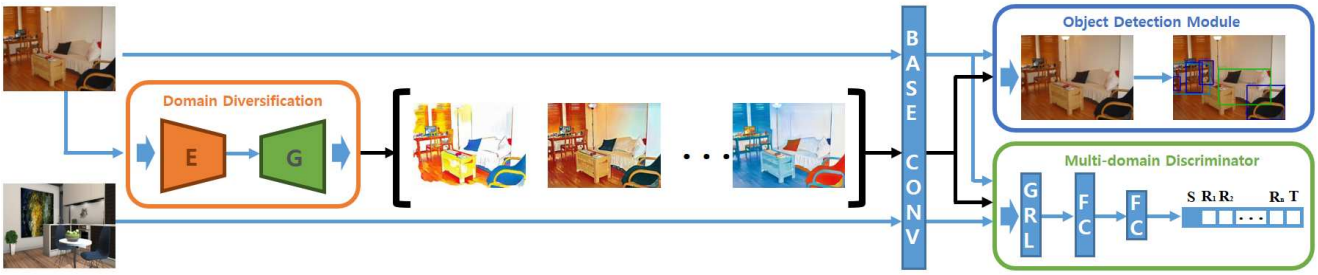


Figure 5. The architecture of our domain adaptation framework for object detection. Our framework is built on the object detection network.

$$L_{\text{MRL}}(x^s, x^t) = L_{\text{mrl}}(G_{\text{Base}}(x^s), 0) + L_{\text{mrl}}(G_{\text{Base}}(x^t), n+1) + \sum_{i=1}^{n} L_{\text{mrl}}(G_{\text{Base}}(G_i(x^s)), i), \quad (6)$$

$$L_{\text{LOC}}(x^s, y^s) = L_{\text{loc}}(x^s, y^s) + \sum_{i=1}^{n} L_{\text{loc}}(G_i(x^s), y^s), \quad (7)$$

$$L_{\text{CLS}}(x^s, y^s) = L_{\text{cls}}(x^s, y^s) + \sum_{i=1}^{n} L_{\text{cls}}(G_i(x^s), y^s), \quad (8)$$

Here, x^s and x^t are images of the source and the target domain, G_{Base} is the base convolutional network in Fig. 5 and y^s is the label information for x^s. In addition, L_{loc} and L_{cls} denote the regression loss and classification loss for the given image, respectively. The overall framework is shown in Fig. 5.

4. Experiments

4.1. Datasets

We verify the effectiveness of our learning paradigm in two different settings: 1) adaptation from real-world to artistic media; 2) adaptation among urban scenes.

Real-world Dataset. PASCAL VOC [10] is a real-world image dataset used for several computer vision tasks. PASCAL VOC 2007 dataset consists of 2,501 train images, 2,510 validation images, and 4,952 test images, while PASCAL VOC 2012 dataset contains 5,717 train images and 5,823 validation images. Annotations are provided for 20 categories. We use train set and validation set on PASCAL VOC 2007 and train set and validation set on PASCAL VOC 2012 as a real-world dataset.

Artistic Media Datasets (AMDs). We use Clipart1k, Watercolor2k, and Comic2k [22] for artistic media domains. These datasets are collected from a website called Behance for the image classification task by [45]. Recently, Inoue et al. [22] notated labels for the object detection task. Each dataset consists of 1,000, 2,000, and 2,000 images, respectively, while half of them are for the test set.

Urban Street Datasets (USDs). We use Cityscapes [7] and Foggy Cityscapes [37] for urban scene datasets. Both of them consist of 2,975 train images and 500 validation images with 8 categories.

Experiment Setup. To validate our method for adaptation tasks from real-world to artistic media, we conduct experiments for Real-world→Clipart1k, Real-world→Watercolor2k, and Real-world→Comic2k. Whole images of each AMD are used for the target domain data during training, while each test set is used for evaluation. For urban scenes, we conduct the experiment for Cityscapes→Foggy Cityscapes. We use Cityscapes train set and Foggy Cityscapes validation set.

4.2. Implementation Details for Domain Shifters

To verify the effectiveness of DD, we generated 3 distinct shifted domains for each adaptation task. Under the universality for domain shifter architecture, we adopt the residual generator and the discriminator from CycleGAN [49]. To distinctively shift the source domain, we consider two factors in the objective function, i.e., color preservation and reconstruction. Figure 6 shows the visual differences caused by each configuration of the factors.

Domain shift considering color preservation: To constraint the domain shifter to preserve color, we adopt the L^1 loss between an input image and a translated image. However, since the instability of the training increases as we give the less effective constraint, we only assign the constraint to the target domain for the diverse shift. Thus, the constraint loss for the domain shifter can be written as:

$$L_{\text{con},1}(G) = \mathbb{E}_{x \sim p_t(x)}[\|(G(x) - x)\|_1]. \quad (9)$$

Domain shift considering reconstruction: To consider the reconstruction, we need one more pair of domain shifter G' and discriminator D' for inverse translation. Moreover, we need additional generative adversarial losses necessary for

Figure 6. Qualitative results for the shifted domains with various configurations of constraint factors. CP and R denote color preservation constraint and reconstruction constraint, respectively.

training G'. Thus, the constraint loss for the domain shifter can be written as:

$$\begin{aligned} L_{\text{con},2}(G, G', D') = & \mathbb{E}_{x \sim p_s(x^s)}[log D'(x^s)] \\ & + \mathbb{E}_{x^t \sim p_t(x^t)}[log(1 - D'(G'(x)))] \\ & + \mathbb{E}_{x^s \sim p_s(x^s)}[\|(G'(G(x^s)) - x^s)\|_1] \\ & + \mathbb{E}_{x^t \sim p_t(x^t)}[\|(G(G'(x^t)) - x^t)\|_1]. \end{aligned} \tag{10}$$

Domain shift considering both reconstruction and color preservation: To consider two factors simultaneously, we apply the sum of two constraint loss terms with additional modules G' and D':

$$L_{\text{con},3}(G, G', D') = L_{\text{con},1}(G) + L_{\text{con},2}(G, G', D'). \tag{11}$$

4.3. Implementation Details for Object Detection

In our experiments, we use Faster R-CNN [35] as our base object detector with VGG-16 [39] pretrained on ImageNet. Each batch consists of $(n + 2)$ images where n is a number of shifted domains. We alleviate the memory issue through gradient accumulation. We train the network for 80k iterations, 50k iteration with a learning rate of 0.001 and the last 30k iterations with a learning rate of 0.0001. All implementations are done in PyTorch and on a single GeForce Titan XP GPU.

For PASCAL VOC and AMDs, we resize the images to have a length of 600 pixels as its shorter side. For USDs, we match the shorter side of the image to be a length of 500 pixels. We evaluate mean average precisions (mAP) in the test phase, following the IoU threshold of 0.5 in [22] and [4]. We follow [35] for unspecified hyper-parameters.

4.4. Performance Comparison

In this section, we compare our method to the state-of-the-art methods (i.e., Domain Adaptive Faster R-CNN (DAF) [3] and Domain Transfer (DT) stage of [22]). For our methods, We apply three shifted domains implemented in section 4.2.

Table 1, 2, 3, and Fig. 7 present the comparison results on Faster R-CNN backbone. Our learning paradigm achieves the highest class-wise AP among all methods in all adaptation tasks except table class in Clipart1k, car class in Watercolor2k. and bus class in Cityscapes. Specifically, for the animal classes in AMDs, our proposed method obtains significantly higher class-wise performance than other methods. To interpret the results in detail, we observe that it is hard to train object detectors with the real-world data to infer discriminatively among animal classes in the artistic media data. However, our learning scheme significantly improves the performance values for the animal classes. Moreover, our method exceeds the state-of-the-art methods by $3\% \sim 12\%$ mAP. Especially for the Real-world $\rightarrow$ AMD tasks, our method outperforms the state-of-the-art methods by around $9\% \sim 12\%$ mAP. These results demonstrate that our method is effective at learning domain-invariant discriminative features and adapting object detection layers to the common feature space, which is further analyzed in section 4.6 and 4.7. Several qualitative results are shown in Fig. 8.

4.5. Ablation Study on Numbers of Shifted Domains

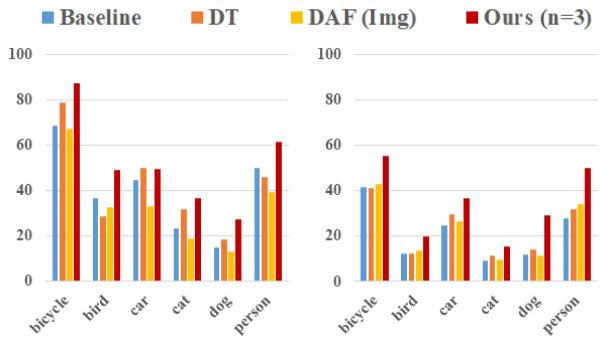
We investigate the effectiveness of the DD stage and the MRL stage on different numbers of the shifted domains. We used the Real-world $\rightarrow$ Clipart1k task as a study case. As shown in Table 4, the overall results of each learning scheme are improved as the number of shifted domains increases. Furthermore, using DD with MRL significantly boosts the performance for overall cases. It is noteworthy that the improvement in performance through MRL is amplified as the number of domains increases. These results validate our hypothesis that DD enhances the domain adaptation effect of the following feature-level adaptation by alleviating the source-biased discriminativity.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Baseline	13.9	51.5	20.4	10.1	29.5	35.1	24.6	3.0	34.7	2.6	25.7	13.3	27.2	47.9	37.5	40.6	4.6	9.1	27.5	40.2	24.9
DT [22]	16.4	62.5	22.8	31.9	44.1	36.3	27.9	0.7	41.9	13.1	**37.6**	5.2	28.0	64.8	58.2	42.7	9.2	19.8	32.8	47.3	32.1
DAF (Img) [4]	20.0	49.9	19.5	17.0	21.2	24.7	20.0	2.0	30.2	10.5	15.4	3.3	25.9	49.3	32.9	23.6	14.3	5.5	30.1	32.0	22.4
Ours (n=3)	**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**

Table 1. Quantitative results for object detection of Clipart1k [22] by adapting from PASCAL VOC [10].

Method	V → Wa	V → Co
Baseline	39.8	21.4
DT [22]	40.0	23.5
DAF (Img) [4]	34.3	23.2
Ours (n=3)	**52.0**	**34.5**

Table 2. Quantitative results for object detection of Watercolor2k [22] and Comic2k [22] by adapting from PASCAL VOC [10]. We denote PASCAL VOC, Watercolor2k, and Comic2k as V, Wa, and Co, respectively.



(a) Watercolor2k (b) Comic2k

Figure 7. Comparison results for the class-wise AP of Watercolor2k test set and Comic2k test set [22].

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Baseline	17.7	24.7	27.2	12.6	14.8	9.1	14.3	23.2	17.9
DT [22]	25.4	39.3	42.4	24.9	40.4	23.1	25.9	30.4	31.5
DAF (Img) [4]	22.9	30.7	39.0	20.1	27.5	17.7	21.4	25.9	25.7
DAF (Ins) [4]	23.6	30.6	38.6	20.8	**40.5**	12.8	17.1	26.1	26.3
DAF (Cons) [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
Ours (n=3)	**30.8**	**40.5**	**44.3**	**27.2**	38.4	**34.5**	**28.4**	**32.2**	**34.6**

Table 3. Quantitative results for object detection of Foggy Cityscapes [37] by adapting from Cityscapes [7].

DD Configuration				DD	DD+MRL	offset
#SD	CP	R	CP + R	mAP		
0				24.9	-	-
1	✓			31.2	32.4	+1.2
2	✓	✓		32.5	37.8	+5.3
3	✓	✓	✓	33.8	41.8	+8.0

Table 4. Results of the ablation study on configuration of the shifted domains. DD and MRL denote domain diversification and multi-domain-invariant representation learning, respectively. The offset denotes the performance improvement of the object detector through MRL. CP, R, CP+R denote the shifted domains trained with color preservation constraint, reconstruction constraint, and both constraints, respectively, and SD denotes shifted domains.

4.6. Study on Alleviation of the Source-biased Discriminativity

To further verify the alleviation of the source-biased discriminativity by DD, we investigate the localization performance of RPN and the classification accuracy of the Fast R-CNN module on the Faster R-CNN baseline. To compare the positive impact of the domain adaptation methods on the localization capability, We compute mean Intersection-over-Union (mIoU) of the best overlaps predicted from RPN for each instance. The classification accuracy is evaluated with the target domain instances. To evaluate the inference capability of the classification layer in the Fast R-CNN module, we provide the ground-truth value for bounding boxes. We conduct the experiments for the Real-world→Clipart1k case.

As shown in Table 5, all domain adaptation methods significantly improve the localization capability of RPN than baseline. However, the domain adaptation methods with DD achieve significanly higher classification accuracy than the methods without DD. Moreover, even though both DAF and MRL are in a frame of feature-level adaptation, the classification results of two methods show considerable gap. These results demonstrate the importance of the discriminative feature when adapting the domains in feature level. Furthermore, we can confirm our demonstration that feature-level adaptation suffers from the source-biased discriminativity and DD is effective at alleviating this issue.

4.7. Error Analysis on Top Ranked Detections

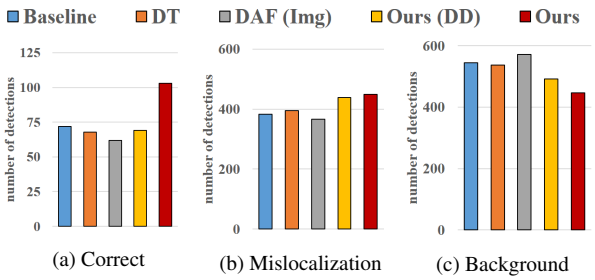
We analyze detection errors to investigate the positive impact of our method on domain adaptation in details. We study Real-world→Clipart1k case for the analysis. Since the Clipart1k test set only has 500 images, we classify the most confident 1,000 detections for each domain adaptation method. With reference to [18], we categorize the detection results into three groups: correct detection, mislocalization error, and background error. Correct detection denotes correct class with IoU greater than 0.5, mislocalization error

(a) Input image (b) Baseline (c) DAF (Img) [3] (d) DT [22] (e) Ours (DD) (f) Ours (DD+MRL) (g) Ground-truth

Figure 8. Qualitative results for object detection of the AMDs by adapting from PASCAL VOC [10]. Images in the first, second, and third rows are from the test sets of Clipart1k, Watercolor2k, and Comic2k [22], respectively. Best view in color.

Method	Acc (%)	mIoU (%)
Baseline	30.6	56.5
DAF (Img)	38.0	65.9
Ours (DD)	50.2	66.6
Ours (DD+MRL)	52.5	68.5

Table 5. Comparison results for the instance classification accuracy of the Fast R-CNN module and mean IoU of RPN for the test set of Clipart1k [22]. Each adaptation method only uses annotations in PASCAL VOC [10].



(a) Correct (b) Mislocalization (c) Background

Figure 9. Error Analysis of the most confident 1,000 detections for each domain adaptation methods.

denotes correct class with IoU between 0.1 and 0.5, and background error denotes wrong class or correct class with IoU less than 0.1, where IoU denotes Intersection-over-Union.

As shown in Fig. 9, both DD with and without MRL reduce background detection errors compared to other methods. However, while both reduce background errors, DD with MRL significantly increases the number of correct detection than DD.

5. Conclusion

In this paper, we have introduced a learning paradigm for object detection to alleviate the chronic limitations of domain adaptation approaches. Our learning paradigm achieves the goal with the incorporation of Domain Diversification (DD) and Multi-domain-invariant Representation Learning (MRL). DD mitigates the source-biased discriminativity of feature-level adaptation by diversifying the distribution of the labeled data. MRL addresses the imperfect image translation by encouraging the unbiased semantic representation among multiple domains. We structurized our learning paradigm into a domain adaptation framework for object detection networks. We confirmed the positive impact of DD and MRL through in-depth analysis, which verifies the effectiveness of our two schemes. Our method outperforms state-of-the-art methods in various cases.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01772, Development of QA systems for Video Story Understanding to pass the Video Turing Test).

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pages 95–104. IEEE Computer Society, 2017.

[2] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. Autodial: Automatic domain alignment layers. In *International Conference on Computer Vision*, 2017.

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. *CoRR*, abs/1803.03243, 2018.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. *CoRR*, abs/1704.08509, 2017.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

[10] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1180–1189. JMLR.org, 2015.

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

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[14] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 222–230, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[16] Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2066–2073, Washington, DC, USA, 2012. IEEE Computer Society.

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[18] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 340–353, Berlin, Heidelberg, 2012. Springer-Verlag.

[19] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Scholkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 601–608, Cambridge, MA, USA, 2006. MIT Press.

[22] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 5967–5976, July 2017.

[24] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865,

International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[25] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[26] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[28] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 469–477. Curran Associates, Inc., 2016.

[29] Songtao Liu, Di Huang, and andYunhong Wang. Receptive field block net for accurate and fast object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[31] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 07–09 Jul 2015. PMLR.

[32] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jia-Guang Sun, Yuchen Guo, and Philip S. Yu. Transfer sparse coding for robust image representation. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 407–414, 2013.

[33] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. Extreme clicking for efficient object annotation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4940–4949, Oct 2017.

[34] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press.

[36] Paolo Russo, Fabio Maria Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *CVPR*, 2018.

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *CoRR*, abs/1708.07819, 2017.

[38] Zhiqiang Shen, Yu-Gang Jiang, Dequan Wang, and Xiangyang Xue. Iterative object and part transfer for fine-grained recognition. In *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*, pages 1470–1475, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI Technical Report, 4th Human Computation Workshop*, 2012.

[41] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, Dec. 2007.

[42] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV 2016 Workshops*, 2016.

[43] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.

[45] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge J. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *CoRR*, abs/1704.08614, 2017.

[46] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. *CoRR*, abs/1711.06897, 2017.

[47] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 00, pages 2039–2049, Oct. 2018.

[48] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.