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

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% ~ 12% in terms of mean average precision (mAP) on various datasets.

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.

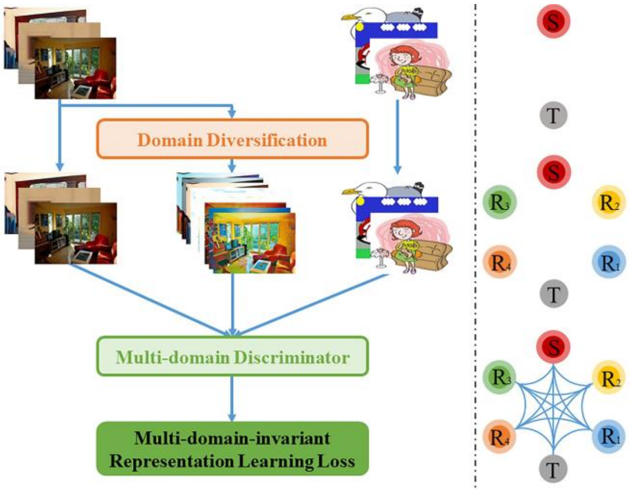

Domain Diversification
Multi-domain Discriminator
Multi-domain-invariant Representation Learning Loss
S
T
S
R3
R2
R4
R1
T
S
R3
R2
R4
R1
T

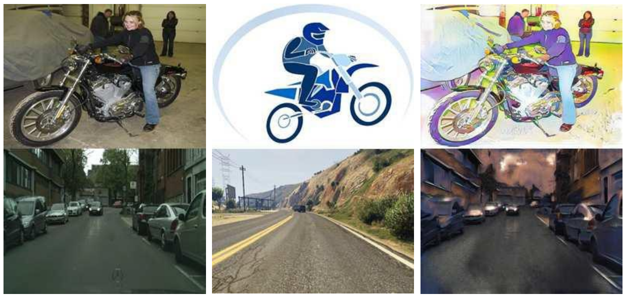
Figure 1. Overview of our learning paradigm. We illustrate a conceptual diagram of the distributions of the domains on the right side. S and T represent for the source and the target domain, respectively, and each R_i represents the ith diversified domain.

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% ~ 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.

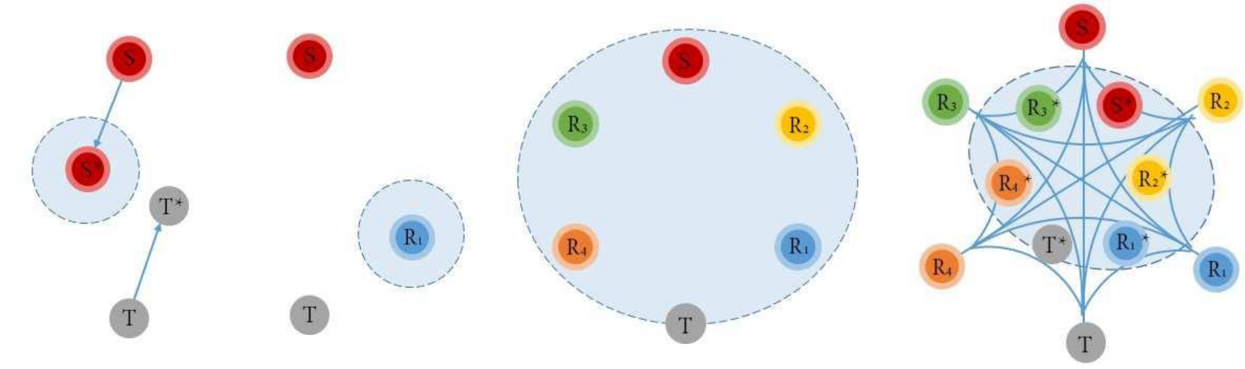
2. Related work

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

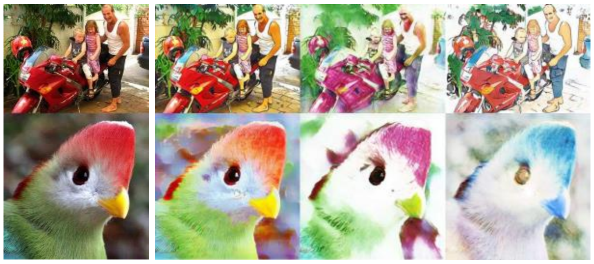
Figure 3. Comparison of distribution transformation by different domain adaptation methods. MRL refers to Multi-domain-invariant Representation Learning. S and T denote the source domain and the target domain, respectively. R_1, R_2, R_3, and R_4 are shifted domains of the source domain. The arrows indicate 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:

$$L_{\mathrm{im}}(G, D, M) = L_{\mathrm{GAN}}(G, D) + aL_{\mathrm{con}}(G, M), \qquad (1)$$

$$\begin{aligned} L_{\mathrm{GAN}}(G, D) = \; & \mathbb{E}_{x^t \sim p_t(x^t)}[logD(x^t)] \\ & + \mathbb{E}_{x^s \sim p_s(x^s)}[log(1 - D(G(x^s)))], \end{aligned} \qquad (2)$$

where D is the discriminator for adversarial learning, $L_{con}(G, M)$ is the constraint loss with a possibly existing additional module M and a 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:

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

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:

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

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:

$$\begin{aligned} \mathrm{L}(x^s, x^t, y^s) = \; & L_{\mathrm{MRL}}(x^s, x^t) + L_{\mathrm{LOC}}(x^s, y^s) \\ & + L_{\mathrm{CLS}}(x^s, y^s), \end{aligned} \qquad (5)$$