AFAN: Augmented Feature Alignment Network for Cross-Domain Object Detection

Hongsong Wang, Shengcai Liao, *Senior Member, IEEE*, and Ling Shao, *Fellow, IEEE*

***Abstract*— Unsupervised domain adaptation for object detection is a challenging problem with many real-world applications. Unfortunately, it has received much less attention than supervised object detection. Models that try to address this task tend to suffer from a shortage of annotated training samples. Moreover, existing methods of feature alignments are not sufficient to learn domain-invariant representations. To address these limitations, we propose a novel augmented feature alignment network (AFAN) which integrates intermediate domain image generation and domain-adversarial training into a unified framework. An intermediate domain image generator is proposed to enhance feature alignments by domain-adversarial training with automatically generated soft domain labels. The synthetic intermediate domain images progressively bridge the domain divergence and augment the annotated source domain training data. A feature pyramid alignment is designed and the corresponding feature discriminator is used to align multi-scale convolutional features of different semantic levels. Last but not least, we introduce a region feature alignment and an instance discriminator to learn domain-invariant features for object proposals. Our approach significantly outperforms the state-of-the-art methods on standard benchmarks for both similar and dissimilar domain adaptations. Further extensive experiments verify the effectiveness of each component and demonstrate that the proposed network can learn domain-invariant representations.**

***Index Terms*— Object detection, unsupervised domain adaptation.**

I. Introduction

OBJECT detection is a fundamental problem in computer vision, and has been extensively studied for decades. Over the past several years, deep learning has achieved remarkable success in this area [1]–[3]. To advance the state-of-the-art on large-scale benchmarks [4], [5], most deep learning based approaches require tremendous amount of annotated training data. In real-world applications, the manual annotating such large-scale data is time-consuming and labor-intensive. Moreover, it is challenging to deploy a trained deep learning model to new environments, even if the task is the same. This is simply because the performance is negatively impacted by changes in conditions such as image sensors, viewpoints, weather and time of day.

There is a non-negligible discrepancy between the distribution of data from the target domain and that from the source domain. Unsupervised domain adaptation [6]–[8] addresses this problem and helps adequately improve the learning in the target domain. Recently, adversarial domain adaptation [9]–[11], which aligns feature distributions between the two domains through adversarial training, has become very popular. Although domain adaptation has achieved great progress in computer vision, most of the studies are restricted to image classification [12], [13] and semantic segmentation [14]–[16]. How to effectively deploy an object detector in a new environment still remains an open problem.

Deep learning based domain adaptive object detection has recently begun to receive attention. Works on this topic can be roughly divided into two categories: feature distribution alignment based methods [17]–[20] and self-training based methods that use pseudo labels [21], [22]. The former approach learns domain-invariant features through adversarial training which uses discriminator networks to predict domain labels for images from the two domains. However, the alignments of convolutional features as well as region features are not sufficient for object detection due to the limited number of annotated images from the source domain. As for the latter approach, the pseudo labels are obtained from the detection model trained only in the source domain. This method requires sophisticated and robust training strategies to overcome the adverse effects of severely noisy labels.

In order to alleviate the above shortcomings, we address cross-domain object detection from a new perspective by introducing an intermediate domain. The intermediate domain is considered as the interpolating path between the source and the target domains. We aim to propose an effective framework which takes advantages of both feature distribution alignment and pseudo image generation by combining intermediate domain image generation and domain-adversarial training.

To this end, we propose a novel augmented feature alignment network (AFAN). An outline of AFAN is illustrated in Figure 1. We introduce an intermediate domain image generator which produces pseudo intermediate domain images. This generator augments the annotated data in the source domain with unlabeled images in the target domain. We theoretically prove that intermediate domain images decrease the divergence in distributions between the two domains. Moreover, we propose a feature pyramid alignment in order to transfer the semantics and reduce the divergence of both high-level and low-level features between different domains. A feature discriminator is designed to align the distributions of convolutional feature maps of multiple scales. Finally, we present an instance discriminator which tackles domain

Manuscript received July 30, 2020; revised January 29, 2021 and March 8, 2021; accepted March 9, 2021. Date of publication April 1, 2021; date of current version April 7, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei. *(Corresponding author: Shengcai Liao.)*

The authors are with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: hongsongsui@gmail.com; scliao@ieee.org; ling.shao@ieee.org).

Digital Object Identifier 10.1109/TIP.2021.3066046

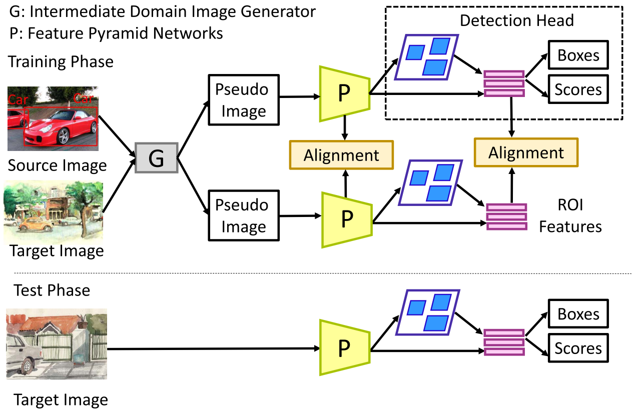


Fig. 1. Outline of the proposed framework. During training, the network behaves like a Siamese neural network which takes two sets of images from the source and target domains, respectively. The test process is a single path which is the same as that of object detection.

shifts for object region proposals. Both the feature discriminator and the instance discriminator are incorporated in the object detection framework through domain-adversarial training. The soft ground-truth domain labels of intermediate domain images enhance the distribution alignment, and the whole network learns domain-invariant representations that cannot be distinguished by either discriminator.

In summary, the main contributions of this paper are as follows:

- We propose an augmented feature alignment network for cross-domain object detection, which integrates intermediate domain image generation and domain-adversarial training into a single framework.
- We present an intermediate domain image generator which creates intermediate domain images between the source and target domains, and theoretically demonstrate that these images reduce the divergence between the two domains.
- We propose the feature pyramid alignment which performs a unified domain adaptation for both high-level and low-level convolutional feature maps.
- Our approach achieves new state-of-the-art performance on various benchmarks of both similar and dissimilar domain adaptations.

The remainder of the paper is organized as follows. Section II reviews related work. Section III details the structure of the proposed method as well as the training method. Comprehensive experimental results, visualizations and ablation studies are presented in Section IV. The conclusions are finally drawn in Section V.

II. Related Work

We briefly review approaches mostly related to ours from three aspects: general object detection, unsupervised domain adaptation and domain adaptation for object detection.

A. Object Detection

Object detection aims to locate and classify the object instances in an input image. Current state-of-the-art approaches can be roughly divided into two categories: one-stage detector and two-stage object detector. Two-stage detectors first predict objectness proposals and then refine the locations and classify the object categories in the second stage. R-CNN [1], Fast R-CNN [2] and Faster R-CNN [3] are milestone works for this pipeline. Faster R-CNN presents a region proposal network (RPN) to generate region proposals, and jointly trains the RPN with the detection network. Influential extending works are Faster R-CNN with Feature Pyramid Network (FPN) [23], Mask R-CNN [24], etc. Other attempts of object detection include learning rotation-invariant features [25], self-supervised feature augmentation [26].

In contrast, one-stage object detectors directly regress the candidate object boxes and classify the object categories in one step. Many approaches, such as include YOLOv2 [27], SSD [28] and RetinaNet [29], use anchor boxes to enumerate possible locations, scales and aspect ratios of objects. Other methods follow anchor-free pipeline which directly learn object possibilities and bounding box coordinates. The representative works in this category include YOLO [30], CSP [31], FoveaBox [32], FCOS [33]. Different from other object objectors which are fine-tuned from the off-the-shelf networks, ScratchDet [34] robustly trains the one-stage object detectors from scratch.

We follow the two-stage pipeline and choose Faster R-CNN as the base detector. Discriminator networks are integrated into the detector, and domain-adversarial training is utilized to learn domain-invariant representations.

B. Unsupervised Domain Adaptation

Domain adaptation aims at learning a model that reduces the distribution shift between an unlabeled target domain and a labeled source domain [6]. Traditional methods bridge this gap by learning a common feature representation across domains [7], [35] or estimating instance weights to reduce sample selection bias [8]. Deep domain adaptation methods embed adaptation modules in deep architectures to learn transferable representations. Yosinski *et al.* [36] discuss the transferability of different layers and demonstrate that the transferability of features decreases as the distance between domains increases. Long *et al.* [37], [38] present deep adaptation network to learn transferrable features by enhancing feature transferability in task-specific layers and matching embedded features in the reproducing kernel Hilbert spaces. Lu *et al.* [39] use the class mean to learn class-specific linear projections for domain adaptation without explicitly modeling the discrepancy between domains.

Inspired by generative adversarial networks (GAN) [40], recent adversarial domain adaptation methods [10], [11] utilize a domain discriminator to distinguish images of the source domain from those of the target domain. This domain discriminator is jointly trained with deep networks which learn representations that are indistinguishable by the discriminator. The adversarial adaptation methods are divided into three types: gradient reversal-based [9], minimax optimization-based [41], and generative adversarial net-based [42]. Ganin and Lempitsky [9] demonstrate that the domain adaptation behavior can be achieved by the gradient reversal layer (GRL). Tzeng *et al.* [41] maximize domain confusion based

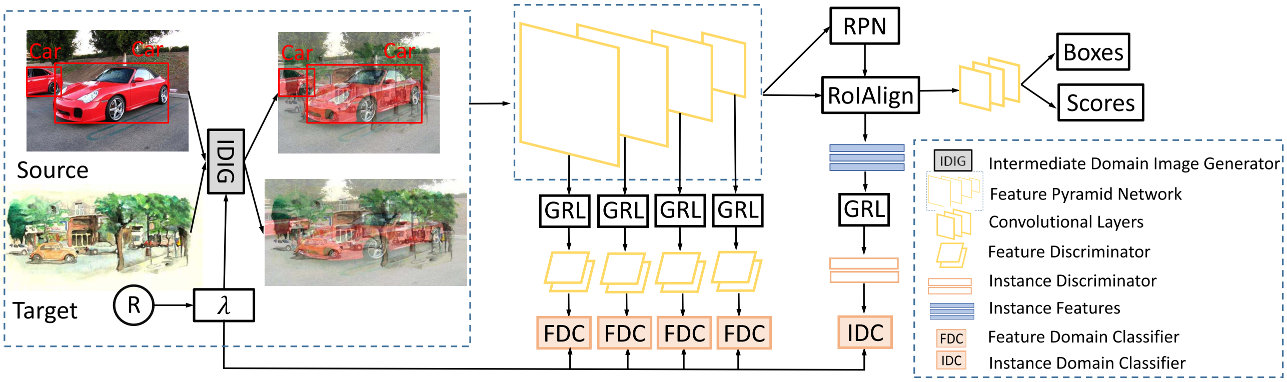


Fig. 2. Pipeline of the proposed AFAN. It is an end-to-end trainable network which consists of four modules: pseudo image generation, feature pyramid alignment, region feature alignment and object detection head. Without loss of generality, the pseudo images are produced by a relatively simple method.

on marginal distributions and transfer correlations between classes from the source domain to the target domain. Hoffman *et al.* [42] propose the cycle-consistent adversarial domain adaptation (CyCADA) to adapt representations in both pixel-level and feature-level without the use of aligned image pairs.

The data augmentation method *mixup* [43] trains a neural network on convex combinations of pairs of examples and their labels for image classification. MixMatch [44] mixes both labeled and unlabeled data with label guesses for semi-supervised classification. Domain mixup [45] mixes between source and target domain images in the adversarial domain adaptation framework for unsupervised domain adaptive classification. However, these methods merely focus image classification, and mixup-based approaches for unsupervised object detection has not yet been explored.

The domain adaptation approaches in computer vision mainly focus on image classification [12], [13] and semantic segmentation [14]–[16]. Our work addresses adversarial domain adaptation for object detection which requires aligning representations of the same object between diverse domains and locating all the related objects in a new domain image.

C. Domain Adaptation for Object Detection

Unsupervised domain adaptation for object detection has recently gained interest [17], [18], [20]–[22], [46]–[50]. Faster R-CNN has been adapted for domain adaptation by aligning the distributions of the last convolutional feature map and the region features [17]. Strong alignment of local features from lower layers and weak alignment of global features from higher layers are explored for convolutional features [18]. Discriminative regions are also mined with a grouping strategy for better alignment for region features [20]. An image-to-image translation via GAN is used to generate images shifted from the source domain to the target domain for pixel-level adaptation [19]. The mean teacher paradigm is applied and the object relation between image regions is used to bridge the domain gap [48]. The attention-based region transfer and prototype-based semantic alignment are proposed to achieve coarse-to-fine feature adaptation [51]. The category-level domain alignment is derived based on graph-based information propagation among features of region proposals [52]. The hierarchical transferability calibration network is introduced to harmonize transferability and discriminability in the context of adversarial adaptation [53] A plug-and-play categorical regularization component is presented to match crucial image regions and important instances across domains [54]. Multiple adversarial domain classifiers are introduced to align the distributions of both local-level features and global-level features [55]. However, current feature alignment methods are still insufficient as different discriminators are applied to different convolutional features. To the best of my knowledge, adversarial domain adaptation with a feature pyramid has not been investigated for cross-domain object detection. The limited amount of annotated training data from the source domain also impedes the learning of domain-invariant representations.

There are also self-training approaches which use trained models in the labeled source domain to generate pseudo labels for unlabeled images from the target domain. However, it is tricky to design a robust learning method to reduce false negatives and false positives. In contrast, we leverage pseudo images for feature alignment which allows us to bypass the problems of previous approaches.

III. Method

Cross-domain object detection aims to guide an object detection model trained on labeled data from a specific source domain to achieve good performance on data from another target domain. The training data consists of labeled data from the source domain $\{(x^s, y^s), \cdots\}$, and unlabeled data from the target domain $\{x^t, \cdots\}$. The detected object classes are assumed to be contained in both domains. Since distributions of image data and object regions from separate domains are different, domain adaptation techniques are required to reduce the discrepancies between domains from the perspectives of images as well as object regions.

A. Framework Overview

We propose a novel augmented feature alignment network (AFAN) which is able to dramatically reduce the divergence across different domains. The pipeline of AFAN is shown in Figure 2. We choose the R-CNN [3] pipeline and adopt deep residual networks [56] as the backbone. We introduce the intermediate domain to bridge the connection between the source and the target domains. An intermediate domain image generator is designed to augment both samples in the source and target domains and enhance feature distribution alignments. A feature pyramid structure is built to transfer information between high- and low-level features. Two diverse discriminator networks, i.e., feature discriminator and instance discriminator, are designed to align distributions of image and object features across the two domains. Both discriminator networks and domain classifiers are incorporated into existing object detection networks. The domain adaptation is accomplished by the proposed discriminator networks through adversarial training.

In the training phase, the proposed network consists of two identical branches which process images from the source and target domains, respectively. As the two branches share parameters, there exist only one branch during testing. The Siamese network structure during training could avoid the imbalanced training data between two domains and make sampling strategy for each dataset more flexible.

B. Intermediate Domain Image Generator

Instead of directly aligning the source and target domains, we introduce the intermediate domain which gradually connects the two domains. The intermediate domain is considered as the interpolation points between the source and target domains. The intermediate domain image generator (IDIG) is an image-to-image module, and both the inputs and outputs are one set of annotated images and another set of unlabeled images. Inspired by *mixup* in image recognition and semi-supervised classification [43], [57], we propose a simple but effective method that generates pseudo training images by interpolating between labeled and unlabeled images. It should be noted that although the recent work [45], [58], [59] also use the mixup strategy to generate pseudo images for unsupervised domain adaptation or adversarial domain adaptation, they merely focus on the task of image classification for which the input image is small and contains a single object. In contrast, we address cross-domain object detection and aim to better align features between diverse domains at different levels, ranging from low-level and high-level convolutional features to regional features.

During training, the IDIG receives two mini-batches of training images from the source and target domains, respectively. The intermediate domain images are generated as

$$\begin{aligned}\tilde{x}^s &= (1-\lambda)x^s + \lambda x^t \\ \tilde{x}^t &= (1-\lambda)x^t + \lambda x^s \end{aligned} \tag{1}$$

where x^s is a labeled image from the source domain, x^t is an unlabeled image from the target domain, $\tilde{x}^s$ and $\tilde{x}^t$ are corresponding pseudo labeled and unlabeled intermediate domain images from the two domains, respectively, and λ is a random variable. As the input images possess various sizes and aspect ratios, during the addition operation, the second image of the formula is resized to the same size of the first one. For each mini-batch, λ is sampled from $\mathbb{U}(0, \lambda_m)$, where λ_m is the upper limit of λ, and $\lambda_m \leq 0.5$.

The IDIG reduces the divergence of distributions between the two domains at the image level. We prove this hypothesis using the generalized energy distance [60] between the distributions of random vectors. Assume that X_s and X_t are random variables of images from the source and target domains with cumulative distribution functions S and T, respectively, the generalized energy distance between S and T is

$$\varepsilon^{(\alpha)}(S,T) = 2\mathbb{E}|X_s - X_t|^\alpha - \mathbb{E}|X_s - {X_s}'|^\alpha - \mathbb{E}|X_t - {X_t}'|^\alpha \tag{2}$$

where X_s and ${X_s}'$ are two independent and identically distributed (iid) random variables from S, X_t and ${X_t}'$ are iid random variables from T, and $0 < \alpha < 2$. From Proposition 2 in [60], when $\alpha = 2$, the distance is reduced as

$$\varepsilon^{(2)}(S,T) = 2|\mathbb{E}(X_s) - \mathbb{E}(X_t)|^2 \tag{3}$$

Let $\tilde{X}_s$ and $\tilde{X}_t$ be the random variables of pseudo intermediate domain images from the source and the target domains, respectively, and $\tilde{S}$ and $\tilde{T}$ be the corresponding cumulative distribution functions, respectively. The generalized energy distance between $\tilde{S}$ and $\tilde{T}$ is

$$\varepsilon^{(2)}(\tilde{S},\tilde{T}) = 2|\mathbb{E}(\tilde{X}_s) - \mathbb{E}(\tilde{X}_t)|^2 \tag{4}$$

In our task, $\tilde{X}_s = (1-\lambda)X_s + \lambda X_t$, $\tilde{X}_t = (1-\lambda)X_t + \lambda X_s$. Thus, the expectation $\mathbb{E}(\tilde{X}_s)$ is computed as

$$\mathbb{E}[(1-\lambda)X_s + \lambda X_t] = (1-\bar{\lambda})\mathbb{E}(X_s) + \bar{\lambda}\mathbb{E}(X_t) \tag{5}$$

where $\bar{\lambda}$ is the expectation of λ. A similar formula can be obtained for $\mathbb{E}(\tilde{X}_t)$.

Therefore, $\varepsilon^{(2)}(\tilde{S},\tilde{T})$ can be written as

$$\varepsilon^{(2)}(\tilde{S},\tilde{T}) = (1-2\bar{\lambda})^2\varepsilon^{(2)}(S,T) \tag{6}$$

Since λ is sampled from $\mathbb{U}(0,\lambda_m)$, $\bar{\lambda} = 0.5\lambda_m$. When $\lambda_m = 0.5$, $\varepsilon^{(2)}(\tilde{S},\tilde{T}) = 0.25\varepsilon^{(2)}(S,T)$, which means that the divergence of the pseudo intermediate domain images between the two domains is much smaller than that of the original images. It should be noted that it is necessary to set an upper bound (e.g., 0.5) on λ during sampling. It is inappropriate to set $\lambda_m > 0.5$ as the labels of the pseudo labeled image $\tilde{x}^s$ would be unreliable with noises of unlabeled images dominating the image content. In addition, $\tilde{x}^s$ would become more similar to x^t instead of x^s, which contradicts the evidence that $\tilde{x}^s$ comes from the source domain. The same analysis applies to $\tilde{x}^t$. As a result, the IDIG separates the large domain gap into small ones, and the augmented images overcome the lack of annotated samples in the source domain. Together with adversarial domain adaptation, the IDIG also enhances the feature distribution alignment, which is discussed below.

It should be noted that the mixup inspired approach is only an example of our IDIG module, and we have provided a theoretical explanation about the benefit of domain adaptation from the energy function perspective. We believe that other image-to-image approaches (e.g., [61]) are also feasible, and investigations about the implementations of the IDIG are beyond the scope of this paper.

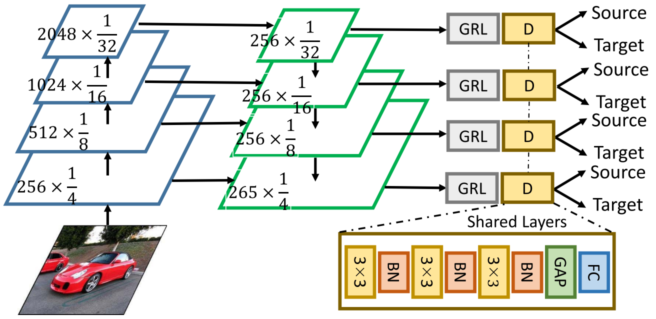


Fig. 3. Structure of feature pyramid alignment. The symbol *D* denotes feature discriminator, which is shared across different convolutional feature maps, and GAP denotes global average pooling.

C. Feature Pyramid Alignment

Aligning the features of deep convolutional neural networks (CNN) between the source and target domains is challenging as there are many different layers in a deep CNN. Some works [17] only align the features of the last convolutional layer, and do not fully bridge the gap across the two domains. Other works [18] use various strategies to align the higher and lower layers. However, such a model is cumbersome as it involves multiple discriminator networks, and it is often difficult to determine whether a CNN layer is high or low.

Inspired by feature pyramid networks (FPN) [23], we propose feature pyramid alignment (FPA) which can align multi-scale features of different layers with only one discriminator network. The structure of FPA is illustrated in Figure 3. The bottom-up pathway generates rich semantical features in the higher layers, and the top-down pathway transfers the semantics from the high layers to the low layers. Due to the lateral connection, the multi-scale convolutional features are transformed in order to have the same feature dimension (numbers of channels). Then, a single convolutional discriminator network is used to classify the domain categories where 0 denotes the source domain and 1 denotes the target domain. The discriminator network is jointly optimized with the object detection networks. While the loss for object detection on labeled data from the source domain is minimized, the loss of the domain classifier for data from both domains should be maximized in order to learn domain-invariant features. During implementation, we adopt the gradient reverse layer (GRL) [9] which leaves the input unchanged during forward propagation and reverses the gradient during back-propagation.

The original ground-truths of domain categories are hard binary labels. However, for the pseudo intermediate domain images generated by the IDIG (see Section III-B), the ground-truths of domain categories are soft labels. The soft labels denote probability distributions between the two domains, which can also be regarded as the weights of images from different domains in the process of intermediate domain image generation. For domain images from the source domain, the domain category label is a two-dimensional vector $[1-\lambda, \lambda]^T$, and for those from the target domain, this label is $[\lambda, 1-\lambda]^T$, where λ is different for each mini-batch during training.

Let F denote the backbone of FPN, and D_f denote the feature discriminator which predicts the probability of the target domain category with respect to convolutional features. The feature discriminator comprises several convolutional layers intertwined with batch normalization layers. A binary domain classifier is placed on top of the discriminator. The unsupervised loss for feature level alignment is

$$\begin{aligned} L_f = -\{&\mathbb{E}[\mu \log(D_f(F(\tilde{X}))] \\ &+ \mathbb{E}[(1-\mu)\log(1-D_f(F(\tilde{X})))]\} \end{aligned} \tag{7}$$

where $\tilde{X} \in \{\tilde{X}_s, \tilde{X}_t\}$ is the pseudo intermediate domain image from a particular domain, and μ is the second component of the soft label for domain categories.

D. Region Feature Alignment

Generating region proposals is an important step for current state-of-the-art object detection approaches. As the input image contains multiple objects, each region proposal accounts for one possible object instance. For cross-domain object detection, adaptation at the proposal level can be attained by aligning the features of region proposals between the two domains. We use RoIAlign [24] to extract features from each proposal, and transform these features with fully connected layers to obtain a 1024-dimensional vector.

As illustrated in Figure 2, an instance discriminator and an instance domain classifier are utilized for adversarial domain adaptation. The instance discriminator network comprises two fully connected layers, and predicts domain labels for individual object instances. The GRL is also used during back-propagation to maximize the loss of the instance domain classifier. Let P denote the features of a region proposal, and D_o denote the instance discriminator which predicts the probability of the target domain category. We assume that the domain label of a region proposal P is the same as that of the image $\tilde{X}$. The loss for proposal level alignment is

$$\begin{aligned} L_o = -\{&\mathbb{E}[\mu \log(D_o(P)] \\ &+ \mathbb{E}[(1-\mu)\log(1-D_o(P))]\} \end{aligned} \tag{8}$$

where μ is λ if the pseudo image is from the source domain and $1-\lambda$ otherwise.

The proposal level adaptation module is applied to intermediate domain images generated by the intermediate domain image generator. Unlike previous approaches [17], [20] which use 0 or 1 as the ground-truth domain labels of real images, our instance level domain classifier exploits soft domain labels as the ground-truths of pseudo images. One of the advantages of intermediate domain images is that the corresponding soft domain labels augment the source domain and bridge the domain divergence in a progressive manner.

E. Training

The object detection loss L_{det} consists of the localization loss and the classification loss. Combined with two types of discriminative losses, the final training loss of the AFAN is

$$L = L_{\text{det}} + \alpha L_f + \beta L_o \tag{9}$$

Algorithm 1 Training Process of the AFAN

```
Input: A batch of labeled source images $\{(x^s, y^s), \cdots\}$ from the source domain, a batch of unlabeled target images $\{x^t, \cdots\}$.
Draw $\lambda$ and $\gamma$ from $\mathbb{U}(0, \lambda_m)$ and $\mathbb{U}(0, 1)$, respectively.
if $\gamma > 0.5$ then
    $\lambda \leftarrow 0$.
end if
Generate intermediate source domain images $\{(\tilde{x}^s, y^s), \cdots\}$ and intermediate target domain images $\{\tilde{x}^t, \cdots\}$ using Equation (1).
Perform the forward pass of deep networks by feeding $\{(\tilde{x}^s, y^s), \cdots\} \cup \{\tilde{x}^t, \cdots\}$.
Calculate the feature level and proposal level discriminator losses with regard to $\lambda$.
Perform the backward pass by minimizing the loss defined in Equation (9).
Output: The updated network.
```

where α, β are weight parameters to balance the detection loss and domain adaptation losses.

During training, the inputs contain two sets of images: labeled images from the source domain and unlabeled images from the target domain. After intermediate domain image generation, the network behaves like a Siamese neural network and computes responses for the two sets of images. During testing, the domain adaptation modules can be discarded and the network takes one image as input.

Details of the training process are summarized in Algorithm 1. In our implementation, an additional parameter γ is introduced to combine both the original images and the pseudo images for training. The explanation of $\lambda < 0.5$ has been discussed in Section III-B. When $\gamma > 0.5, \lambda = 0$, the generated pseudo images are the same as the original images.

IV. Experiments

The proposed approach is evaluated under different experimental settings, and compared with previous state-of-the-art methods. Ablation studies and qualitative experiments are also provided.

A. Experimental Setup

The experimental settings of cross-domain object detection can be divided into two types: adaptation between similar domains and adaptation between dissimilar domains.

1) Adaptation Between Similar Domains: This setting includes adapting normal images to images under different weather and daytime conditions. We use the Cityscapes [62] dataset and Foggy Cityscapes [63] dataset as the source and target domains, respectively. Both datasets have 2,975 images in the training set, and 500 images in the validation set. Foggy Cityscapes is a synthetic foggy dataset and the images simulate fog in real scenes. Following the split of [17], we use annotated images from the training set of Cityscapes and only images from the training set of Foggy Cityscapes for training. The results are evaluated on the validation set of Foggy Cityscapes.

TABLE I

Results of Cross-Domain Pedestrian Detection on the EuroCity Persons Dataset. The Domain Is Transferred From Day Condition to Night Condition

Method	AP of pedestrian
FPN [23]	73.4
Baseline	73.4
Oracle	84.1
Ours	**78.5**

As object detection at night is challenging and night images are difficult to annotate, we adapt object detection from day images to night images. We exploit cross-domain pedestrian detection and evaluate the proposed method on the EuroCity Persons [64] dataset. EuroCity Persons is a large-scale dataset with over 238,000 persons instances manually labeled in over 47,000 images of urban traffic scenes. This dataset has subsets of both day and night, and each subset is split into training and validation sets. For the day subset, the numbers of images in the training set and the validation set are 23892 and 4266, respectively. For the night subset, these numbers are 4222 and 770, respectively. We consider the day subset as the source domain and the night subset as the target domain. The labeled day training set and unlabeled night training set are used for training, and the night validation set is used for validation. To the best of our knowledge, this is one of the first studies on cross-domain pedestrian detection.

2) Adaptation Between Dissimilar Domains: For this setting, images from the two domains are collected under very different scenarios. For example, the source data include synthetic images captured from video games, while the target data are realistic images. The SIM 10k [65] dataset is treated as the source domain, and the Cityscapes dataset is the target domain. SIM 10k contains 10,000 synthetic images with bounding boxes of cars. For the target domain, only the images in the training set are used for training and the results are evaluated on the Cityscapes validation set.

Two datasets collected by different devices in different environments also constitute very dissimilar domains. We regard the Cityscapes dataset as the source domain and KITTI dataset [66], [67] as the target domain. As the categories of the two datasets are a bit different, we classify person sitting and pedestrian as person, van and car as car, tram as train, cyclist as rider in the KITTI dataset. Our setting is different from [17] which is limited as it only considers the category of car. Since this dataset does not have standard splits, all the training images are used for both training and validation.

Compared with realistic photographs, it is more challenging to detect objects in artistic images which embody a breadth of styles, media, and emotions. Cross-domain object detection is adapted from the real-world Pascal VOC [68] to the artistic Watercolor [47]. The Watercolor dataset contains images posted by professional and commercial artists, and has six classes in common with classes of the Pascal VOC. It includes 2000 images with 1000 images for the training set and 1000 images for the test set. In accordance with the setup of [47], training and validation sets of both VOC2007 and VOC2012 are considered as the source domain, and the Watercolor is used as the target domain.

TABLE II

RESULTS OF CROSS-DOMAIN OBJECT DETECTION ADAPTED FROM THE CITYSCAPES TO THE FOGGY-CITYSCAPES

Method	person	rider	car	truck	bus	train	motor	bicycle	mAP
DA-Faster [17]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
DT [47]	25.4	39.3	42.4	24.9	40.4	23.1	25.9	30.4	31.5
S-Align [20]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
SW-Align [18]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
DD-MRL [19]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
MTOR [48]	30.6	41.4	44.0	21.9	38.6	**40.6**	28.3	35.6	35.1
RLDA [22]	35.1	42.1	49.2	**30.1**	45.2	27.0	26.8	36.0	36.4
SW-Faster-ICR-CCR [54]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
ART-PSA [51]	34.0	**46.9**	52.1	30.8	43.2	29.9	34.7	37.4	38.6
PSA [51]	33.5	45.2	51.5	28.2	41.6	26.6	**36.9**	35.4	37.4
Baseline	32.6	35.3	37.1	17.6	28.5	7.3	21.9	28.2	26.1
Oracle	46.6	47.8	64.9	31.2	47.7	48.4	32.4	40.7	45.0
Ours	**42.5**	44.6	**57.0**	26.4	**48.0**	28.3	33.2	**37.1**	**39.6**

3) Implementation Details: To build a feature pyramid, ResNet-50 [56] is adopted as the backbone network due to its simplicity and efficiency. FPN [23] constructs four levels of features with different spatial scales. The feature discriminator consists of three convolutional layers with a 3×3 convolutional kernel. Unless otherwise specified, the channel numbers of both the input and the output are 256. A global pooling layer is placed before the fully connected layer with two neurons that classifies the domain categories. The instance discriminator consists of two fully connected layers which first reduce the dimension of the region features to 512 and then conduct domain classification. For each image, we select the top 1000 proposals with confidence scores above 0.05 for region feature alignment. The domain classifiers are trained with the binary cross-entropy loss.

The whole network is trained using stochastic gradient descent (SGD) with a momentum of 0.9. The base learning rate is 0.01, and the batch size and the training epoch are 16 and 120, respectively. We use four GPUs for training. During testing, the minimum confidence score threshold is 0.05, and NMS is used as post-processing. For all experiments, we evaluate the detection results using mean average precision (mAP) with an IoU threshold of 0.5.

Many previous approaches such as [17], [51], [54] adopt VGG16 [70] as the backbone which does not involve the FPN. As the recent CNN architectures (e.g., ResNet) have shown substantial performance improvement over the VGG, we use ResNet-50 with FPN as the backbone, which is more appropriate for practical applications. The implementation is based on Mask-RCNN benchmark [69].

B. Similar Domains Adaptation

The results of cross-domain pedestrian detection are shown in Table I. As we are the first to perform unsupervised pedestrian detection at night by transferring knowledge from the day domain, there is no reported results of previous approaches on this benchmark. For the EuroCity Persons, *baseline* denotes Faster R-CNN which uses the training set of the day subset for training. Since the training set of the night subset has a much smaller number of images compared to that of the day subset, we consider Faster R-CNN that uses all annotated training images from the two subsets for training as the oracle upper limit, which is denoted as *oracle*. Both *baseline* and *oracle* adopt the same backbone and have the same settings as our approach. We observe that the proposed AFAN outperforms the *baseline* by 5.1%, which clearly demonstrates the effectiveness of the proposed domain adaptation in pedestrian detection from day to night images.

TABLE III

CAR DETECTION ADAPTED FROM THE SIM 10K TO THE CITYSCAPES

Method	AP of car
DA-Faster [17]	39.0
SW-Align [18]	42.3
S-Align [20]	43.0
ART-PSA [51]	43.8
Baseline	32.9
Oracle	68.6
Ours	**45.5**

The results of object detection adapted from Cityscapes to Foggy-Cityscapes are summarized in Table II. We compare the average precision for each category as well as the mAP. Similarly, *baseline* denotes Faster R-CNN trained on the annotated Cityscapes training set. *Oracle* is the Faster R-CNN method trained on the annotated Foggy-Cityscapes training set. In other words, *Baseline* is the method without domain adaptation, and *oracle* is the upper limit. The proposed AFAN outperforms all the recent methods, and achieves an absolute improvement of 3.2% over the best detector reported on this dataset. It also achieves the state-of-the-art performance for most classes. For the averaged performance, our AFAN outperforms *baseline* by 13.5%, which is significant as the margin between AFAN and *oracle* is only 5.4%.

C. Dissimilar Domains Adaptation

In Table III, we compare the results of object detection adapted from the synthetic images to the real images. *Baseline* and *oracle* denote the Faster R-CNN method which trained on the training sets of the SIM 10k and the Cityscapes, respectively. The proposed AFAN considerably outperforms the recent state-of-the-art approaches. In particular, the average precision of AFAN is 2.5% higher than that of the recent method [20], and 12.6% higher than that of *baseline*.

The results of adaptation from the Cityscapes dataset to the KITTI dataset are shown in Table IV. As both the training and validation processes share the same unlabeled images from the target domain, we do not present the results of *oracle*. Our approach consistently shows dramatic improvement over

TABLE IV

RESULTS OF CROSS-DOMAIN OBJECT DETECTION ADAPTED FROM THE CITYSCAPES TO THE KITTI

Method	person	rider	car	truck	train	mAP
DA-Faster [17]	40.9	16.1	70.3	23.6	21.2	34.4
PSA [51]	50.2	27.3	73.2	29.5	17.1	39.5
ART-PSA [51]	50.4	**29.7**	73.6	**29.7**	21.6	41.0
Baseline	54.9	15.7	71.9	31.8	20.6	38.9
Ours	**57.7**	18.5	**74.7**	28.4	**27.6**	**41.4**

TABLE V

RESULTS OF CROSS-DOMAIN OBJECT DETECTION ADAPTED FROM THE PASCAL VOC TO THE WATERCOLOR

Method	bike	bird	car	cat	dog	person	mAP
DA-Faster [17]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
DT [47]	82.8	47.0	40.2	34.6	35.3	62.5	50.4
WST-BSR [21]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
Baseline	80.2	39.8	45.5	28.3	18.3	46.8	43.1
Oracle	86.5	56.4	51.4	39.9	42.3	74.7	58.5
Ours	87.0	46.4	47.3	33.1	30.0	60.1	**50.6**

TABLE VI

ABLATION STUDY RESULTS FOR THE PROPOSED METHOD. FOR SIMPLICITY, THE INTERMEDIATE DOMAIN IMAGE GENERATOR, FEATURE PYRAMID ALIGNMENT, AND REGION FEATURE ALIGNMENT ARE ABBREVIATED AS IDIG, FPA AND RFA, RESPECTIVELY. THE SYMBOL CITYSCAPES→FOGGY DENOTES MAP ADAPTED FROM THE CITYSCAPES TO THE FOGGY-CITYSCAPES, SIM→CITYSCAPES DENOTES CAR DETECTION ADAPTED FROM THE SIM 10K TO THE CITYSCAPES, AND DAY→NIGHT DENOTES PEDESTRIAN DETECTION ADAPTED FROM THE EUROCITY PERSONS DAY SUBSET TO THE EUROCITY PERSONS NIGHT SUBSET

Method	Cityscapes→Foggy	SIM→Cityscapes	Day→Night
AFAN	39.6	45.5	78.5
AFAN w/o RFA	38.4	44.1	77.8
AFAN w/o FPA	31.3	37.2	75.7
AFAN w/o IDIG	34.8	38.1	73.4

baseline and yields the state-of-the-art performance. For example, for the train category, the proposed AFAN outperforms *baseline* by nearly 7.0%. These experiments demonstrate the effectiveness of our approach for cross-domain object detection even if the two domains are very different.

The results of on the artistic Watercolor dataset are provided in Table V. Similar to other experiments, our approach significantly improves the performance of the *baseline* without domain adaptation, and achieves new state-of-the-art performance for the mean average precision on the artistic media dataset.

D. Experimental Analysis

We conduct extensive experiments to investigate the effect of the proposed discriminators and intermediate domain image generator for cross-domain object detection.

1) Ablation Studies: We run a number of ablations to analyze the proposed model. The results are summarized in Table VI. We find that, without the intermediate domain image generator or feature pyramid alignment, the results decrease dramatically. When removing one of the two modules, the mAP on the Foggy Cityscapes decreases 4.8% and 8.3%, respectively, and the AP of car on the Cityscapes decreases 7.4% and 8.3%, respectively. Without region feature alignment, the results also reduce considerably. For example, The mAP on the Foggy Cityscapes drops 1.2% without region feature alignment. These experiments verify the effectiveness of the three modules, which are very complementary to each other. The three components are strongly connected and integrated into a single network during training.

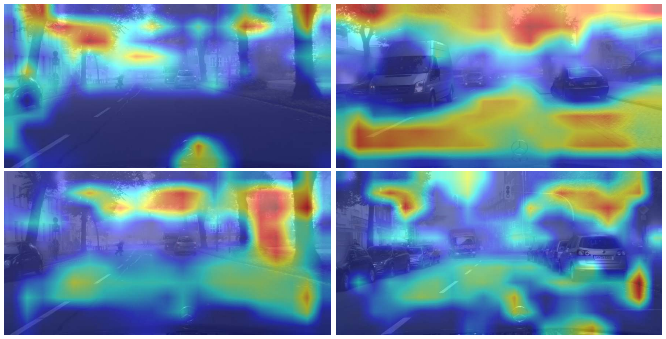


Fig. 4. Evidence of domain classifiers in images from Foggy Cityscapes. The gradient of the domain classification loss is propagated backwards and important regions are highlighted in the image using Grad-CAM [71]. The first and second rows show examples from the source and target domains, respectively.

2) Visualization of Domain Evidence: To investigate the roles of discriminator networks and domain classifiers in network training, we visualize the evidence of the feature discriminator in Figure 4. Similar results can also be obtained for the instance discriminator. The heatmap images highlight the regions where the domain classifier thinks the image comes from the source or the target. For images both domains, the domain classifier does not look at objects such as cars and persons. Instead, the background regions are highlighted and considered as evidence by the feature discriminator. This indicates that the network seems to focus on objects to deceive the domain classifier. In other words, the network demonstrates the ability to learn domain-invariant representations of objects.

3) Visualization of Features: To demonstrate that the proposed AFAN learns domain-invariant features and that the intermediate domain image generator (IDIG) enhances feature distribution alignments, we visualize learned features in Figure 5. We take the domain adaptation experiment from Cityscapes to Foggy Cityscapes as an example. To obtain one representation for each image, we average across the feature maps for the convolutional features and also average all the region features. The features are represented by two-dimensional points after dimensionality reduction to guarantee that similar features are represented by nearby points and dissimilar features are represented by distant points with high probability.

From the figures, we can see that, for the *baseline*, features of the source domain are distant from those of the target domain without domain adaptation. Without the IDIG module, features from the target are only partially aligned with those from the source. However, the learned features from the two domains are distributed closely for the proposed AFAN. The results indicate that AFAN w/o IDIG reduces the feature distribution divergence between the two domains to

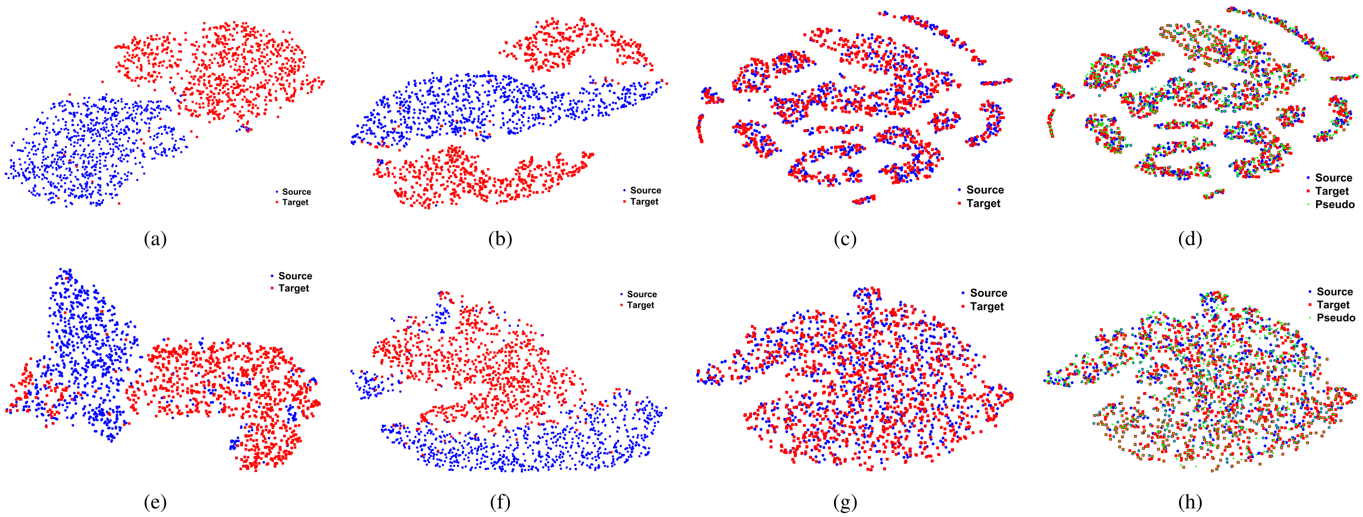


Fig. 5. Visualization of features learned from images. Cityscape and Foggy Cityscape are regarded as the source and the target domains, respectively. The first two columns show features of images learned by *baseline* and the proposed AFAN w/o IDIG, respectively. The third column illustrates features of images extracted by our AFAN. Images from the source and the target domains are indicated by blue dots and red squares, respectively. The last column shows features of intermediate domain images generated by the IDIG module. The first and second rows show convolutional features of the backbone and regional features of object proposals, respectively. For rich visualizations, the feature dimensionality is reduced to two by t-SNE. The more similar feature distributions between the two domains, the better object detection results.

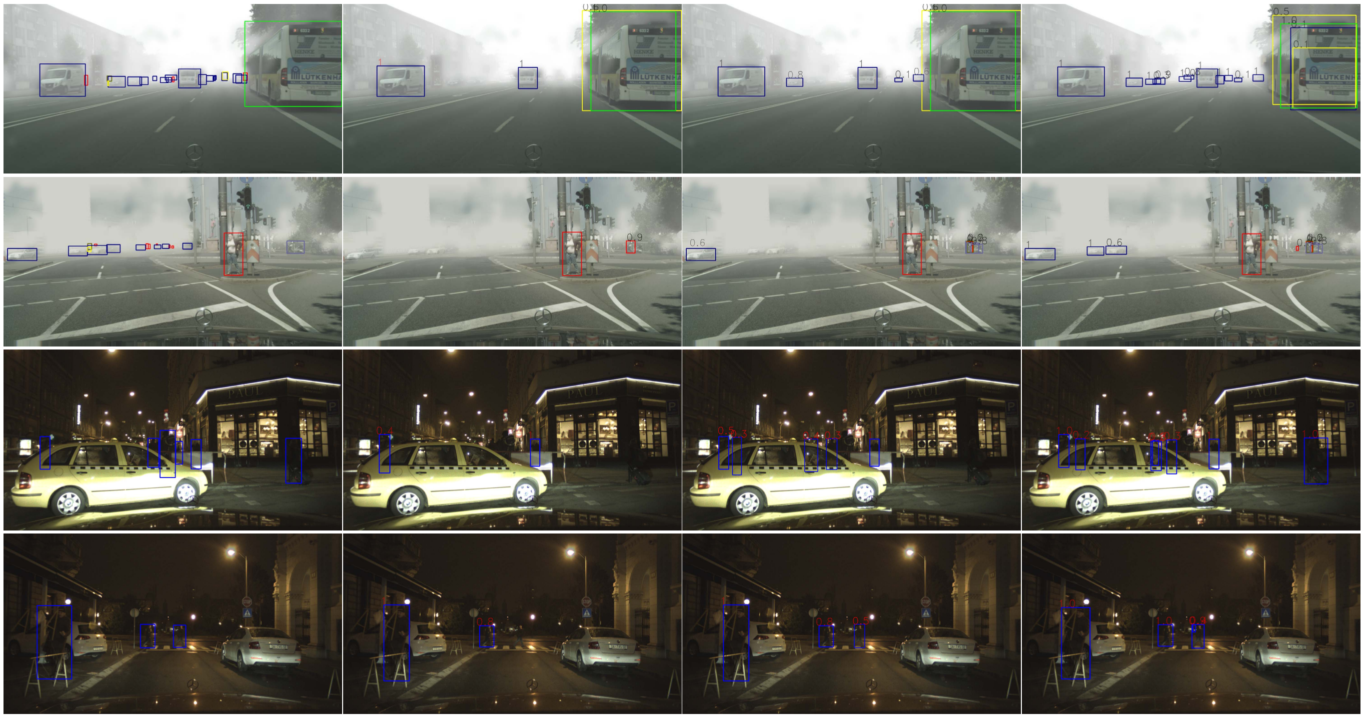


Fig. 6. Examples of detection results on the target domain. From left to right, the four columns correspond to the ground truth, results of Faster R-CNN [3], Domain Adaptive Faster R-CNN [17], and the proposed AFAN, respectively. The first and second rows show images from the validation set of Foggy Cityscapes. The third and fourth rows show images from the validation set of the night subset of EuroCity Persons. Boxes of different classes are marked with different colors, and the predicted confidence scores are described in the text above the corresponding boxes.

some extent. In contrast, this feature distribution divergence can be almost removed by AFAN. Similar results are obtained for both convolutional and regional features.

We also find that features of pseudo images are no different from those of the original images. It can be interpreted that the pseudo images bridge samples from different domains and facilitate the learning of features that are visually indistinguishable. These experiments highlight the important role of the IDIG module for feature distribution alignment.

4) Examples of Detection Results: The cross-domain object detection results are visualized in Figure 6. As the environmental conditions change, the baseline method usually misses some true positive objects which can be reliably detected by our approach. For example, in the second row, cars in the fog cannot be detected by other methods, while our AFAN can detect them with high confidence scores. Our approach gains similar advantages for pedestrian detection at night.

V. Conclusion

In this work, we attempt to address unsupervised domain adaptation for object detection by progressively bridging the domain divergence with adversarial domain adaptation in an intermediate domain. We propose a novel augmented feature alignment network (AFAN) which integrates an intermediate domain image generator and adversarial feature alignments into a single object detection framework. Our method significantly outperforms state-of-the-art approaches on five datasets for both similar and dissimilar domain adaptations. Ablation studies verify the effectiveness and complementarities of the intermediate domain image generation and adversarial feature alignments. Further experiments indicate the evidence of domain discriminators and reveal the role of enhancing feature alignment of the intermediate domain image generator for cross-domain object detection.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[2] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[5] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[7] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[8] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 601–608.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[10] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[11] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[12] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[13] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 597–613.

[14] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[15] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2020–2030.

[16] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.

[17] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[18] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6956–6965.

[19] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12456–12465.

[20] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 687–696.

[21] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6092–6101.

[22] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 480–490.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[25] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[26] X. Pan *et al.*, "Self-supervised feature augmentation for large image object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6745–6758, 2020.

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[28] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[31] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[32] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[33] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[34] R. Zhu *et al.*, "ScratchDet: Training single-shot object detectors from scratch," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2268–2277.

[35] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, "Domain invariant and class discriminative feature learning for visual domain adaptation," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4260–4273, Sep. 2018.

[36] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[37] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[38] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

[39] H. Lu, C. Shen, Z. Cao, Y. Xiao, and A. van den Hengel, "An embarrassingly simple approach to visual domain adaptation," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3403–3417, Jul. 2018.

[40] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[41] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

[42] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[43] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "*mixup*: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13. [Online]. Available: http://www.OpenReview.net

[44] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.

[45] M. Xu *et al.*, "Adversarial domain adaptation with domain mixup," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 6502–6509.

[46] F. Mirrashed, V. I. Morariu, B. Siddiquie, R. S. Feris, and L. S. Davis, "Domain adaptive object detection," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 323–330.

[47] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5001–5009.

[48] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11457–11466.

[49] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7289–7298.

[50] A. RoyChowdhury *et al.*, "Automatic adaptation of object detectors to new domains using self-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 780–790.

[51] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13766–13775.

[52] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12355–12364.

[53] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8869–8878.

[54] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11724–11733.

[55] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang, "Multi-level domain adaptive learning for cross-domain detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3213–3219.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] Q. Wang, W. Li, and L. Van Gool, "Semi-supervised learning by augmented distribution alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1466–1475.

[58] X. Mao, Y. Ma, Z. Yang, Y. Chen, and Q. Li, "Virtual mixup training for unsupervised domain adaptation," 2019, *arXiv:1905.04215*. [Online]. Available: http://arxiv.org/abs/1905.04215

[59] S. Yan, H. Song, N. Li, L. Zou, and L. Ren, "Improve unsupervised domain adaptation with mixup training," 2020, *arXiv:2001.00677*. [Online]. Available: http://arxiv.org/abs/2001.00677

[60] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *J. Stat. Planning Inference*, vol. 143, no. 8, pp. 1249–1272, Aug. 2013.

[61] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4066–4079, Aug. 2018.

[62] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[63] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.

[64] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "EuroCity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.

[65] 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?" in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2017, pp. 1–8.

[66] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

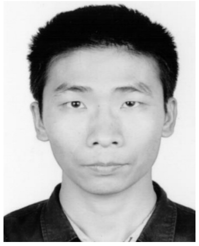
[67] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[68] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[69] F. Massa and R. Girshick. (2018). *MaskrCNN-Benchmark: Fast, Modular Reference Implementation of Instance Segmentation and Object Detection Algorithms in PyTorch*. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

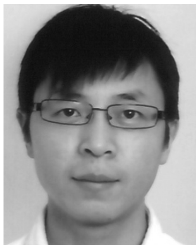
[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



Hongsong Wang received the B.E. degree in automation from the Huazhong University of Science and Technology, in 2013, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, University of Chinese Academy of Sciences, in 2018. He is currently a Researcher with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include deep learning-based applications, such as pedestrian detection, video representation, and action recognition.

Shengcai Liao (Senior Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Sun Yat-sen University, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2010. He was an Associate Professor with CASIA. From 2010 to 2012, he was a Postdoctoral Fellow with the Department of Computer Science and Engineering, Michigan State University. He is currently a Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He has published over 100 articles, with over 13000 citations according to Google Scholar. His research interests include computer vision and pattern recognition, with a focus on image and video analysis, particularly face recognition, object detection, person re-identification, and video surveillance. He was a recipient of the Best Student Paper Award in ICB 2006, ICB 2015, and CCBR 2016, the Best Paper Award in ICB 2007, and the Best Reviewer Award in IJCB 2014 and CVPR 2019 Outstanding Reviewers. He served as an Area Chair for ICPR 2016, ICB 2016, and ICB 2018, and as a PC member for ICCV, CVPR, and ECCV.



Ling Shao (Fellow, IEEE) is currently the CEO and the Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a Fellow of the IAPR, the IET, and the BCS.