OA-FSUI2IT: A Novel Few-Shot Cross Domain Object Detection Framework with Object-Aware Few-Shot Unsupervised Image-to-Image Translation

Lifan Zhao*, **Yunlong Meng***, **Lin Xu**†

AI Institute, Shanghai Em-Data Technology Co., Ltd., China
{zlf.charlie, yunlong.cuhk, lin.xu5470}@gmail.com

Abstract

Unsupervised image-to-image (UI2I) translation methods aim to learn a mapping between different visual domains with well-preserved content and consistent structure. It has been proven that the generated images are quite useful for enhancing the performance of computer vision tasks like object detection in a different domain with distribution discrepancies. Current methods require large amounts of images in both source and target domains for successful translation. However, data collection and annotations in many scenarios are infeasible or even impossible. In this paper, we propose an **O**bject-**A**ware **F**ew-**S**hot **UI2I T**ranslation (OA-FSUI2IT) framework to address the few-shot cross domain (FSCD) object detection task with limited unlabeled images in the target domain. To this end, we first introduce a discriminator augmentation (DA) module into the OA-FSUI2IT framework for successful few-shot UI2I translation. Then, we present a patch pyramid contrastive learning (PPCL) strategy to further improve the quality of the generated images. Last, we propose a self-supervised content-consistency (SSCC) loss to enforce the content-consistency in the translation. We implement extensive experiments to demonstrate the effectiveness of our OA-FSUI2IT framework for FSCD object detection and achieve state-of-the-art performance on the benchmarks of *Normal-to-Foggy*, *Day-to-Night*, and *Cross-scene adaptation*. The source code of our proposed method is also available at https://github.com/emdata-ailab/FSCD-Det.

Introduction

Object detection is a fundamental computer vision task. Recent advances of deep learning (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016) and large amounts of the annotated data (Cordts et al. 2016; Sakaridis, Dai, and Van Gool 2018; Geiger, Lenz, and Urtasun 2012; Yu et al. 2020) propel the fast development of object detection and reach remarkable achievements. However, detection models trained on the source domain may degenerate the performance in a new target domain, due to the domain shift problem (Gopalan, Li, and Chellappa 2011; Patel et al. 2015; Tzeng et al. 2017; Ganin et al. 2016; Doersch and Zisserman 2017) caused by the variances of object appearance, viewpoints, backgrounds, illumination, and weather conditions, etc. In addition, collecting large-scale and diverse datasets with accurate bounding boxes annotations is difficult or even impossible since the labeling process is not only expensive but also time-consuming, and data collection processes in some scenarios are infeasible (Yu et al. 2019; Wang et al. 2019; Zhuang et al. 2020). Serving as an effective solution to bridge the gap of the data distributions between different domains, unsupervised domain adaptation (UDA) (Ganin and Lempitsky 2015; Ganin et al. 2016; Wilson and Cook 2020) is proposed to learn invariant representations explicitly. The learned knowledge is transferred from the train data domain (source domain) to the test data domain (target domain). Usually, the unsupervised domain adaptation can be classified into three main categories: i) statistics matching; ii) adversarial pixel and feature level adaptation; iii) content-preserving image-to-image translation based data augmentation. In the first category, distributional variations across different domains are mitigated via high-order statistics matching of source and target features with well-designed distribution divergence metrics, i.e., maximum mean discrepancy (MMD) (Tzeng et al. 2014; Long et al. 2015), second-order moment (Sun and Saenko 2016), central moment discrepancy (CMD) (Zellinger et al. 2017). Methods belonged to the second category integrate domain adversarial training into the *de facto* detector, e.g. Faster R-CNN (Ren et al. 2015), YOLO (Redmon et al. 2016; Redmon and Farhadi 2017, 2018), SSD (Liu et al. 2016), Mask R-CNN (He et al. 2017). Typically, they attempt to minimize the domain disparity and reach appealing transferability via feature-level and pixel-level alignment (Chen et al. 2018). Methods of third class are built on a common sense that recently emerged conditional genera-

Figure 1: Few-shot UI2I translation and FSCD object detection results with *Normal-to-Foggy* and *Day-to-Night*.

*These authors contributed equally.
†Corresponding author.

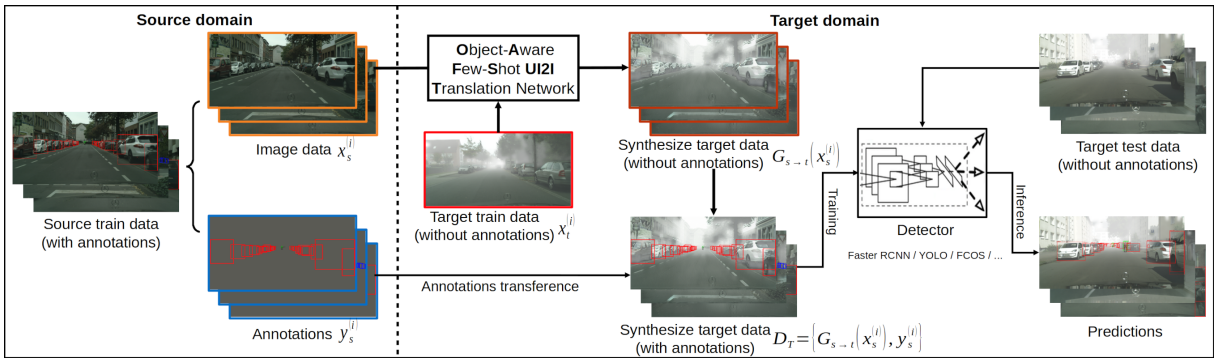

Figure 2: Overview of the FSCD object detection framework.

tive models (van den Oord et al. 2016; Kingma and Welling 2014; Mirza and Osindero 2014) are effective on image style transfer (Jing et al. 2020) with promising content preservation for data augmentation. Often, they first use unsupervised image-to-image (UI2I) translation techniques (Pang et al. 2021; Liu et al. 2019; Saito, Saenko, and Liu 2020; Chen et al. 2020b; Kim et al. 2019a; Liu, Breuel, and Kautz 2017) to generate images from the source domain (have both large-scale images and annotated labels) to the target domain (only have images but no labels). Next, by transferring the annotations to form the new training dataset in the target domain, developed detection models are capable of training, validation, and testing. For example, Arruda *et al.* employ CycleGAN (Zhu et al. 2017) to produce nighttime images with the transferred annotated labels as the training dataset in the night domain (Arruda et al. 2019). Then both two-stage and one-stage detectors, i.e., Faster RCNN and YOLO are capable of model training. Evidently, the methods in the third class can also earn benefits from the continuously raising of the universal object detection performance, compared to the previous two categories. However, current UDA-based cross-domain object detection methods can only improve the detection performance with large amounts of data in the target domain. Significant performance degradations are observed for the limited data case. Although collecting more data in the target domain may alleviate the impact of this data-scarce issue, it is non-trivial as effective data acquisitions in some scenarios are infeasible or even impossible. For example, the raindrops splattered on the camera lens may corrupt the images with blurring and scattering vision effects. Consequently, very few usable rainy images can be captured in practice. Hence, imbalance data amounts between the source and target domains may further hinder learning the valuable features and cause UI2I translation and domain adversarial training failures.

In this paper, we solve the few-shot cross-domain (FSCD) object detection task by proposing a novel **O**bject-**A**ware **F**ew-**S**hot **UI2I** **T**ranslation (OA-FSUI2IT) network for content-preserving few-shot UI2I translation and use the synthesized images of OA-FSUI2IT network as the training dataset in the target domain for off-the-shelf universal detector training to boost the FSCD detection performance. Our OA-FSUI2IT based FSCD object detection framework consists of two primary components: 1) OA-FSUI2IT network, and 2) *de facto* detector. First, in the OA-FSUI2IT network, we propose discriminator augmentation (DA) in the target domain to resolve the data imbalance issue and style inadequate problem for successful few-shot UI2I translation. We present the patch pyramid contrastive learning (PPCL) strategy for coherent associations at each specific location and propose self-supervised content consistent (SSCC) loss to further enforce the content preservation performance. Later, for a detector, by transferring the annotations from the source images to the corresponding generated images in the target domain, we can formulate the synthesized labeled training dataset and directly benefit FSCD object detection. To the best of our knowledge, the proposed OA-FSUI2IT based FSCD object detection framework is the first work to successfully resolve the cross-domain detection task in the scenario that only limited unlabeled image data (e.g., only 10 available images) are available in the target domain. We implement extensive experiments to demonstrate the effectiveness of OA-FSUI2IT and achieve state-of-the-art performance on multiple FSCD object detection benchmarks. For example, we reach 43.5 mAP on *Normal-to-Foggy* (*Cityscapes* $\rightarrow$ *FoggyCityscapes*), 30.5 mAP on *Day-to-Night* (*BDD100k Daytime Clear* $\rightarrow$ *BDD100k Nighttime Clear*), 26.3 mAP on *Cross scene adaption* (*Kitti* $\rightarrow$ *Cityscapes*). In a nutshell, the main contributions of this paper are summarized as follows:

1. We propose a novel OA-FSUI2IT framework to address the FSCD object detection task. Up to our best knowledge, this is the first work to address the cross domain detection task successfully with limited unlabeled images in the target domain (e.g., only with 1 or 10 available images).

2. We propose a series of new modules include DA, PPCL, and SSCC into the OA-FSUI2IT framework. These modules work together to alleviate the data imbalance issue and preserve content well for improving FSCD object detection in the target domain.

3. We implement extensive experiments and achieve state-of-the-art performance on multiple FSCD object detection benchmarks (*Cityscapes* $\rightarrow$ *FoggyCityscapes*, *BDD100k Daytime Clear*$\rightarrow$ *Nighttime Clear*, *Kitti* $\rightarrow$ *Cityscapes*), demonstrating the effectiveness of OA-FSUI2IT framework.

Methods

Problem Formulation

We address the FSCD object detection task with the supposition that we have sufficient source labeled data $\mathcal{D}_{\mathcal{S}} = \{(x_s^{(i)}, y_s^{(i)})\}_{i=1}^{N_s}$, but inadequate unlabeled data $\mathcal{D}_{\mathcal{T}} = \{x_t^{(i)}\}_{i=1}^{N_t}$ in the target domain, i.e., $N_s \gg N_t$. $x_s^{(i)}$ and $x_t^{(i)}$ denote the i-th image in the source domain and target domain, respectively. $y_s^{(i)} = (b_s^{(i)}, c_s^{(i)})$ represents the annotations for the i-th source image, $x_s^{(i)}$, where $b_s^{(i)}$ and $c_s^{(i)}$ are the coordinates of bounding box b and its associated category c, respectively. With only a few images but unknown labels in the target domain, our aim is to generate well content preserved images in the target domain with the learned OA-FSUI2IT network and train the detector with the synthesized image dataset after bounding box annotations transference to achieve promising performance on the test image data in the target domain. Fig. 2 presents the processing overview of our OA-FSUI2IT framework for FSCD object detection.

It is comprised of two main parts: OA-FSUI2IT network and detector. The OA-FSUI2IT network is trained to map a "content" image, $x_s^{(i)}$, from the source domain to generate an analogous image, $x_{s\to t}^{(i)} = G_{s\to t}(x_s^{(i)})$, in the target domain, with the input "style" image, $x_t^{(i)}$. Note that our OA-FSUI2IT is operated in an unsupervised setting without pairing constraints between the source and target domains. Our goal is to learn the style from the target input image, $x_t^{(i)}$, but preserve the content for the source input image, $x_s^{(i)}$.

Though only one word different from "Few-Shot Object Detection (FSD)", FSD focuses on detecting novel-class instances, which is completely unrelated to FSCD.

Preliminary Knowledge

CUT Baseline For Few-Shot UI2I Translation. CUT (Park et al. 2020) is built upon the noise contrastive estimation framework (Oord, Li, and Vinyals 2018), uses InfoNCE loss (Gutmann and Hyvärinen 2010) for mutual information maximization for conditional image synthesis, and therefore is capable of associating the input and output data. To simplify the training procedures, reduce the training time and release the cycle-consistency loss constraint, the generator $G_{s\to t}$ in CUT is intentionally designed as two sequential components, an encoder $G_{s\to t}^{enc}$ and a decoder $G_{s\to t}^{dec}$. As for the i-th input image from the source domain, $x_s^{(i)}$, CUT can produce the output image in the target domain as $x_{s\to t}^{(i)} = G_{s\to t}(x_s^{(i)}) = G_{s\to t}^{dec}(G_{s\to t}^{enc}(x_s))$. Then, a multi-layer patch-based learning objective is used for content and feature matching in the specific location of the corresponding input-output patch.

Limitations of CUT. The translation results of CUT suffer from unseen objects in the datasets (Park et al. 2020). This degradation may be further exaggerated when using CUT for few-shot UI2I translation tasks with multiple different scale objects in the scene, as more varieties of the features need to be extracted. Limited data in the target domain would lead to inadequate encoder training and inappropriate feature extraction of the contrastive loss. It further deteriorates the content-preserving capability, and brings about unexpected blurring effects in the translated images, as shown in Fig. 3(a) and 3(b). Usually, blurring issues may cause detection performance degradation. In addition, the translation results of CUT present a serious inadequate style transfer problem. For example, the CUT attempts to generate erroneous lighting effects in many improper places. As shown in Fig. 3(c) and 3(d), windows and corners of the buildings are falsely lightened, which are very different from the realistic images in the target domain. Such deviations come from the insufficient training images (e.g., 10 images in this working) in the target domain. We propose OA-FSUI2IT network, with the inspiration of StyleGAN2-Ada (Karras et al. 2020), DiffAugment (Zhao et al. 2020), and self-supervised learning (SSL), to alleviate such limitations that appears in the CUT.

(a) (b) (c) (d)

Figure 3: Limitations of CUT for few-shot UI2I translation.

OA-FSUI2IT Network

Fig. 4 presents the overall architecture of the OA-FSUI2IT.

Discriminator Augmentation (DA). GAN relies on the discriminator to produce realistic images in UI2I translation task. Therefore, without sufficient data in both the source and target domain, the discriminator may suffer serious overfitting problem and consequently devastate the generator. We design a discriminator augmentation module in our OA-FSUI2IT network to resolve the data imbalance and discriminator overfitting issues, as shown in Fig. 4. During the training process, we apply augment operations, i.e., T, including *pixel blitting* (horizontal flip, rotation, integer translation), *geometric transformation* (isotropic scale, arbitrary rotation, anisotropic scaling, fractional translation), and *color transforms* (brightness, contrast, luma flip, hue, saturation), on the real input target image, i.e., x_t, and the synthesized image in the target domain, i.e., $G_{s\to t}(x_s)$, for the discriminator D_t.[1] Because the style cannot be altered in the augmentation implementing, the augmented images, i.e., $T(x_t)$, and $T(G_{s\to t}(x_s))$, should own the same styles as their original images, i.e., x_t, and $G_{s\to t}(x_s)$. The adversarial loss in our OA-FSUI2IT network with discriminator augmentation module can be expressed as:

$$\begin{aligned}\mathcal{L}_{D_{da}} =& \mathbb{E}_{x_t\sim Y}[f_{D_t}(-D_t(T(x_t)))] \\ &+ \mathbb{E}_{x_s\sim X}[f_{D_t}(D_t(T(G_{s\to t}(x_s))))],\end{aligned} \tag{1}$$

$$\mathcal{L}_{G_{da}} = \mathbb{E}_{x_s\sim X}[f_{G_{s\to t}}(-D_t(T(G_{s\to t}(x_s))))], \tag{2}$$

where generator $G_{s\to t}$ tries to produce images that have similar appearance to the images belonged to the target domain $\mathcal{Y}_{\mathcal{T}}$ and preserve the primary contents with input images

[1]For abbreviation, we use x_s, x_t instead of $x_s^{(i)}$, $x_t^{(i)}$ to denote the image later in this working.

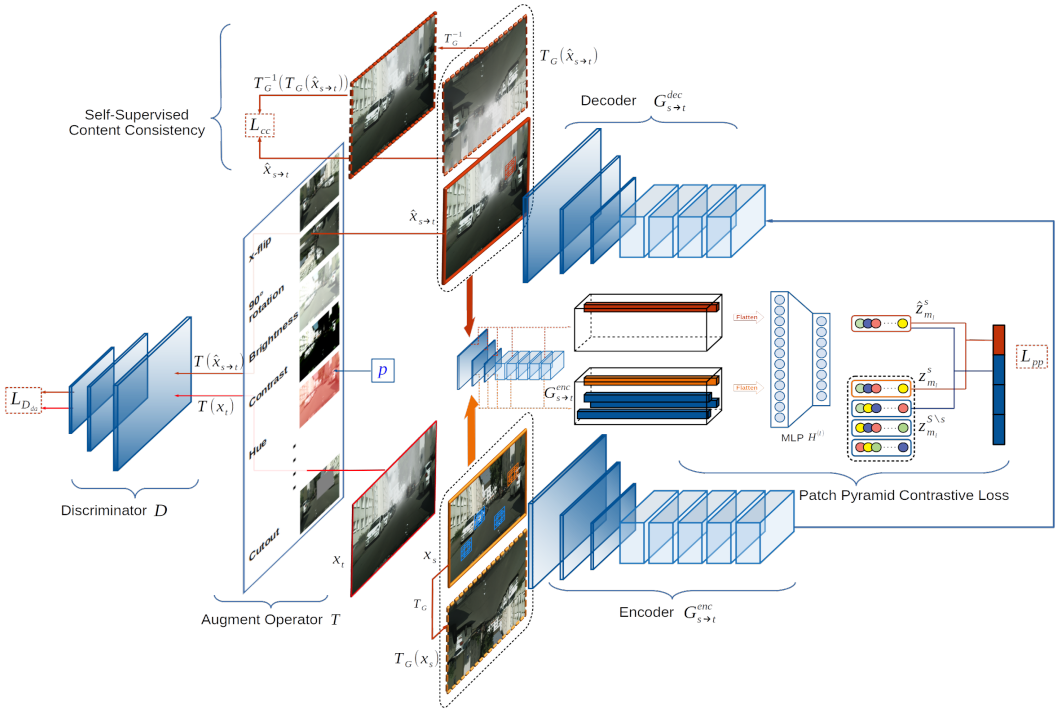

Figure 4: Architecture of the OA-FSUI2IT network. (1) For an input source image x_s, we apply an revertible augmentation transform and get $T_G(x_s)$. Also we acquire a set of pyramid patches $P_m(x_s)$ of different sizes and aspect ratios on this image. (2) Then passing them through the encoder $G^{enc}_{s\to t}$ and decoder $G^{dec}_{s\to t}$ of our generator outputs corresponding $\hat{x}_{s\to t}$ and $T_G(\hat{x}_{s\to t})$. (3) Then we pairwise pass $\{x_s, P_m(x_s)\}$ and corresponding $\{\hat{x}_{s\to t}, G_{s\to t}(P_m(x_s))\}$ through the encoder again to get feature maps and calculate L_{pp} between them. (4) What's more, L_{cc} is obtained via the L1 Loss between $\hat{x}_{s\to t}$ and its reverted transformed image $T_G^{-1}(T_G(\hat{x}_{s\to t}))$. (5) On the discriminator side, we augment the target image x_t and the generated image $\hat{x}_{s\to t}$ with the augment operator T controlled by the possibility p , feed them to the discriminator and get GAN loss $L_{D_{da}}$.

from the source domain $\mathcal{X}_\mathcal{S}$, while discriminator D_t aims to distinguish the translated samples $G_{s\to t}(x_s)$ and the real samples x_t; f_{D_t} and $f_{G_{s\to t}}$ are the given loss functions for the discriminator D_t and generator $G_{s\to t}$.

Patch Pyramid Contrastive Learning (PPCL). CUT employs a multilayer, patch-based contrastive learning objective for content sharing on both image and patch levels. However, it fails to ensure the coherent associations for the extracted nearby patches. In addition, CUT only uses a single patch size in the training process. Even the objects belong to the same class usually present multiple different scales. Hence, vallina CUT is not suitable to generate images with multiscale objects and may significantly degrade the performance of the object detector. We propose patch pyramid contrastive learning to resolve this issue. Similar to the anchor boxes used in Faster R-CNN (Ren et al. 2015) and default boxes used in SSD (Liu et al. 2016), we associate a set of pyramid patches of different sizes and aspect ratios at a specific location, illustrated in Fig. 4. We employ *PatchPyramidNCE* loss to match the corresponding input-output patch pair for each single patch, and the summation of pyramid patches formulate our *patch pyramid contrastive loss*($\mathcal{L}_{pp}$), expressed as:

$$\mathcal{L}_{pp} = \mathbb{E}_{x_s\sim\mathcal{X}_\mathcal{S}} \sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{s=1}^{S_{m_l}} \rho(\boldsymbol{z}^s_{m_l}, \hat{\boldsymbol{z}}^s_{m_l}, \boldsymbol{z}^{S\setminus s}_{m_l}), \tag{3}$$

where M is the number of patch scales in the pyramid set; L is the number of selected layers of interest; $s \in \{1, ..., S_{m_l}\}$, S_{m_l} is the number of selected spatial locations for the l-th layer of interest in the m-th patch scale; $\{\boldsymbol{z}^s_{m_l}\}^{m_L}_{m_l=m_1}$ is a stack of features in the m-th patch scale, which is produced as: $\boldsymbol{z}^s_{m_l} = H^{(l)}(G^{enc}_{s\to t}{}^{(l)}(P_m(x_s)))$; $P_m(x_s)$ is the m-th patch for the input source image x_s; $H^{(l)}$ is a two-layer MLP network with 256 hidden neurons; $G^{enc}_{s\to t}{}^{(l)}$ is the output feature map of the l-th layer of interest; $\rho(\boldsymbol{z}^s_{m_l}, \hat{\boldsymbol{z}}^s_{m_l}, \boldsymbol{z}^{S\setminus s}_{m_l})$ represents the probability of selecting the corresponding "positive" patch, $\boldsymbol{z}^s_{m_l}$, over the non-matching "negative" patches, $\boldsymbol{z}^{S\setminus s}_{m_l}$, for the query, $\boldsymbol{z}^s_{m_l}$, given by [2]:

$$\rho(\boldsymbol{z}^s_{m_l}, \hat{\boldsymbol{z}}^s_{m_l}, \boldsymbol{z}^{S\setminus s}_{m_l}) = -\log\left[\frac{\exp(\frac{\boldsymbol{z}^s_{m_l}\cdot\hat{\boldsymbol{z}}^s_{m_l}}{\tau})}{\exp(\frac{\boldsymbol{z}^s_{m_l}\cdot\hat{\boldsymbol{z}}^s_{m_l}}{\tau}) + \sum_{n=1,n\neq s}^{S_{m_l}}\exp(\frac{\boldsymbol{z}^s_{m_l}\cdot\boldsymbol{z}^{S\setminus s}_{m_l}}{\tau})}\right] \tag{4}$$

[2]Here, the cross-entropy loss is employed to associate the input and output patch pyramid. We set temperature $\tau = 0.07$.

Method	Detector	Backbone	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	gain
Source Only	Faster RCNN	VGG16	24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3	-
Source Only	YOLOv3	DarkNet-53	29.5	38.3	41.4	18.0	28.9	5.0	22.9	33.8	27.2	-
Source Only	FCOS	ResNet50-FPN	21.9	17.0	29.9	2.8	11.5	5.0	6.1	25.1	14.9	-
Source Only	Faster RCNN	ResNet50-FPN	38.6	45.1	44.9	18.8	24.6	2.8	23.5	42.0	30.0	-
***Ten-Shot** Target*												
DDMRL	Faster RCNN	VGG16	27.6	38.1	42.9	17.1	27.6	14.3	14.6	32.8	26.9	+6.6
SWDA	Faster RCNN	VGG16	25.5	30.8	40.4	21.1	26.1	**34.5**	6.1	13.4	24.7	+4.4
CycleGAN	Faster RCNN	ResNet50-FPN	30.7	38.5	57.6	17.9	31.4	4.6	9.4	34.9	28.1	-1.9
DCLGAN	Faster RCNN	ResNet50-FPN	39.7	50.3	56.7	18.9	34.1	13.9	22.7	45.4	35.2	+5.2
CUT	YOLOv3	DarkNet-53	33.9	43.2	51.9	23.2	40.3	17.8	27.2	36.4	34.3	+7.1
CUT	FCOS	ResNet50-FPN	33.1	30.6	48.3	7.6	23.6	1.7	6.4	34.9	23.3	+8.4
CUT	Faster RCNN	ResNet50-FPN	44.5	53.4	58.3	22.8	36.6	14.9	34.4	51.3	39.5	+9.5
OA-FSUI2IT	YOLOv3	DarkNet-53	34.9	42.7	56.6	21.2	**48.1**	32.9	20.9	37.1	36.8	+9.6
OA-FSUI2IT	FCOS	ResNet50-FPN	36.7	35.4	56.2	11.3	25.3	1.7	12.8	35.0	26.8	+11.9
OA-FSUI2IT	Faster RCNN	ResNet50-FPN	**47.5**	**53.8**	**64.1**	**27.8**	45.9	11.5	**35.9**	**52.3**	**42.3**	**+12.3**
***One-Shot** Target*												
DAFRCNN	Faster RCNN	VGG16	30.4	36.3	41.4	18.5	32.8	9.1	20.3	25.9	26.8	+6.5
OSHOT	Faster RCNN	ResNet50	32.1	46.1	43.1	20.4	39.8	15.9	27.1	32.4	31.9	+1.9
CUT	Faster RCNN	ResNet50-FPN	39.8	47.2	53.7	24.5	34.7	11.0	22.1	44.0	34.6	+4.6
OA-FSUI2IT	Faster RCNN	ResNet50-FPN	**44.5**	**52.2**	**58.8**	**25.6**	**44.2**	**22.3**	**30.4**	**48.3**	**40.8**	**+10.8**
Oracle	Faster RCNN	VGG16	36.2	47.7	53.0	34.7	51.9	41.0	36.8	37.8	42.4	+22.1
Oracle	YOLOv3	DarkNet-53	38.2	47.5	60.6	30.9	47.7	36.8	35.4	40.3	42.2	+15.0
Oracle	FCOS	ResNet50-FPN	40.5	38.8	61.1	14.9	33.0	11.0	8.4	36.2	30.5	+15.6
Oracle	Faster RCNN	ResNet50-FPN	53.1	59.9	71.1	31.3	48.5	29.1	39.8	55.8	48.6	+18.6

Table 1: Detection results on *Normal-to-Foggy* (*Cityscapes* $\rightarrow$ *FoggyCityscapes*). Average precision (%) is reported in the target domain. Note that we cite the quantitative results of *DAFRCNN* (Chen et al. 2018), *OSHOT* (Innocente et al. 2020), *DDMRL* (Kim et al. 2019b), and *SWDA* (Saito et al. 2019) from (Innocente et al. 2020), while *Source Only* and *Oracle (VGG16 backbone)* from (Chen et al. 2020a). The best results of *Ten-Shot* and *One-Shot* are bolded.

Here, we can refer $\boldsymbol{z}^s_{m_l} \in \mathbb{R}^{C_{m_l}}$ as the "query" feature; $\hat{\boldsymbol{z}}^s_{m_l}$ is the corresponding "positive" patch, in the encoded output image, as $\hat{\boldsymbol{z}}^s_{m_l} = H^{(l)}(G^{enc}_{s\rightarrow t}{}^{(l)}(P_m(G_{s\rightarrow t}(x_s))))$; $\boldsymbol{z}^{S\setminus s}_{m_l} \in \mathbb{R}^{(S_{m_l}-1)\times C_{m_l}}$ is the non-matching "negative" patches; C_{m_l} is the number of channels for the l-th layer in the m-th patch.

Self-Supervised Content Consistency (SSCC) Loss. We propose to derive supervision signals from the image data itself to further enforce content consistency for our OA-FSUI2IT network, with the inspiration of recent advances in SSL regime. We assume that applying an augment transform pair, (T_G, T_G^{-1}), on the input end and output end, respectively, should not change the synthesized result, i.e., $G_{s\rightarrow t}(x_s) = T_G^{-1}(G_{s\rightarrow t}(T_G(x_s))$. Here, T_G is an reversible augment operation, i.e., flipping, translation, and linear scaling, and T_G^{-1} is its inverse transform. We employ such provenance information to formulate the proxy task to further enforce the content consistency in the translation. We use L_1 norm of the two generated images, one without augmentation and another with augment transform, as shown in Fig. 4, to formulate self-supervised content consistency loss ($\mathcal{L}_{cc}$):

$$\mathcal{L}_{cc} = \frac{1}{N_{T_G}} \sum_{k=0}^{N_{T_G}} [G_{s\rightarrow t}(x_s) - T_G^{-1}{}^{(k)}(G_{s\rightarrow t}(T_G^{(k)}(x_s)))] \quad (5)$$

where $T_G^{(k)}$ and $T_G^{-1}{}^{(k)}$ are the k-th augmenting way, and its inverse transform, respectively; $G_{s\rightarrow t}$ is the mapping function; x_s is the input image in the source domain, $\mathcal{D}_\mathcal{S}$; $G_{s\rightarrow t}(x_s)$ is the generated image in the target domain, $\mathcal{D}_\mathcal{T}$; N_{T_G} is the number of augment transformation; and $T_G^{-1}{}^{(k)}(G_{s\rightarrow t}(T_G^{(k)}(x_s)))$ is the generated image in the target domain, $\mathcal{D}_\mathcal{T}$, with the k-th augment transform pair.

Method	Detector	person	rider	car	train	mAP	gain
Source Only	FCOS	22.8	18.9	38.8	0.6	20.3	-
Source Only	Faster RCNN	25.3	18.2	37.5	0.1	20.3	-
CycleGAN	Faster RCNN	11.2	3.8	31.3	6.0	13.1	-7.2
DCLGAN	Faster RCNN	19.6	7.5	36.0	16.7	20.0	-0.3
CUT	FCOS	18.9	12.0	41.3	1.9	18.5	-1.8
CUT	Faster RCNN	22.6	12.6	36.1	15.4	21.7	+1.4
OA-FSUI2IT	FCOS	24.3	16.4	**42.7**	4.2	21.9	+1.6
OA-FSUI2IT	Faster RCNN	**27.5**	**16.9**	42.3	**18.7**	**26.3**	**+6.0**
Oracle	FCOS	49.0	52.6	66.7	9.0	44.3	+24.0
Oracle	Faster RCNN	57.4	63.4	74.0	24.5	54.8	+34.5

Table 2: Detection performance comparison results on *Cross scene adaptation* (*Kitti* $\rightarrow$ *Cityscapes*). Average precision (%) is reported in the target domain.

Full Objective. Finally, we can formulate full objective as:

$$\mathcal{L} = \mathcal{L}_{D_{da}} + \lambda_{pp}\mathcal{L}_{pp} + \lambda_{cc}\mathcal{L}_{cc} \quad (6)$$

where λ_{pp} and λ_{cc} are the trade-off factors for the patch pyramid contrastive learning and the self-supervised content-consistent loss, respectively. The generated images should be realistic, while patches in the input and output images should share the correspondence.

Experiments

We present implementation details of the OA-FSUI2IT framework for the task of FSCD object detection, and demonstrate its effectiveness on multiple benchmarks. The training data for OA-FSUI2IT network consists of: i) the

Figure 5: Translation results for *Normal-to-Foggy* task.

Figure 6: Translation results for *Day-to-Night* task.

Figure 7: Translation results for *Cross scene adaptation* task.

source training data with sufficient amounts of images and annotations (bounding boxes and object categories), i.e., N_s is large enough, and ii) the *target training data* with limited number of unlabeled images, i.e., N_t is very small. We set $N_t = 10$ throughout the paper. To validate the performance of our OA-FSUI2IT for all few-shot domain shift scenarios, we report the final results of our model as well as the results by combining different modules. To the best of our knowledge, this is the first work to address the cross domain detection with limited unlabeled images in the target domain.

Datasets and Evaluation

Datasets. We evaluate our OA-FSUI2IT framework for FSCD detection task under three scenarios: i) ***Normal-to-Foggy*** (*Cityscapes* $\rightarrow$ *FoggyCityscapes*), ii) ***Day-to-Night*** (*BDD100k Daytime Clear* $\rightarrow$ *BDD100k Nighttime Clear*), and iii) ***Cross scene adaptation*** (*Kitti* $\rightarrow$ *Cityscapes*).

Evaluation Metrics. We use Fréchet Inception Distance (FID) (Heusel et al. 2017), precision & recall (Sajjadi et al. 2018), and density & coverage (Naeem et al. 2020), to quantitatively measure the image quality for the synthesized images. We evaluate the detection performance by reporting the mean average precisions (mAP) with a threshold of 0.5 for all experiments, as (Chen et al. 2018).

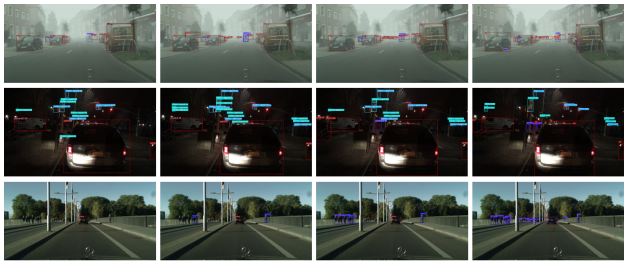

Figure 8: Illustration of the detection results on the target domain. First row: *Normal-to-Foggy*. Second row: *Day-to-Night*. Third row: *Cross scene adaptation*.

Implementation Details

To demonstrate the effectiveness of the three proposed modules, we intentionally match the architecture and hyperparameter settings with CUT (Park et al. 2020). We use the generator with 9 residual blocks (Johnson, Alahi, and Fei-Fei 2016), PatchGAN discriminator (Isola et al. 2017), LSGAN loss (Mao et al. 2017), batch size of 1, and Adam optimizer (Kingma and Ba 2014) with learning rate of 0.002. The parameters λ_{pp} and λ_{cc} are both set to 1.

Method	Detector	Backbone	Bike	Bus	Car	Motor	Person	Rider	Light	Sign	Truck	mAP	gain
Source Only	YOLOv3	DarkNet53	20.9	24.7	51.0	8.6	29.9	16.2	19.3	42.2	31.2	24.4	-
Source Only	FCOS	ResNet50-FPN	9.7	10.4	44.4	1.4	24.9	5.9	13.3	40.3	21.2	17.1	-
Source Only	Faster RCNN	ResNet50-FPN	22.0	21.7	53.0	9.5	33.9	15.3	12.8	40.6	29.3	26.4	-
CycleGAN	Faster RCNN	ResNet50-FPN	20.5	25.5	53.2	6.9	36.1	10.6	10.1	39.5	30.9	25.9	-0.5
DCLGAN	Faster RCNN	ResNet50-FPN	22.4	29.0	47.4	8.1	35.8	12.5	12.8	41.9	26.0	26.2	-0.2
CUT	YOLOv3	DarkNet53	10.6	23.9	47.4	5.9	24.1	9.1	8.0	31.1	22.6	18.3	-6.1
CUT	FCOS	ResNet50-FPN	5.4	9.1	42.7	0.5	21.4	5.5	6.7	34.0	17.2	14.3	-2.8
CUT	Faster RCNN	ResNet50-FPN	22.6	27.2	52.2	4.5	36.3	12.5	10.0	42.3	29.4	26.3	-0.1
OA-FSUI2IT	YOLOv3	DarkNet53	**28.6**	**30.3**	**53.4**	8.3	32.1	20.4	**17.9**	39.5	**33.4**	28.6	+4.2
OA-FSUI2IT	FCOS	ResNet50-FPN	17.4	17.1	49.7	5.0	34.1	10.3	15.6	43.1	26.1	24.3	**+7.2**
OA-FSUI2IT	Faster RCNN	ResNet50-FPN	27.5	28.2	53.3	**16.2**	**39.7**	**20.7**	13.0	**43.7**	32.1	**30.5**	+4.1
Oracle	YOLOv3	DarkNet53	37.0	40.3	67.2	34.7	45.3	30.2	46.0	59.6	49.1	40.9	+16.5
Oracle	FCOS	ResNet50-FPN	27.6	45.6	73.6	29.2	52.5	20.9	57.9	67.0	51.1	42.5	+25.4
Oracle	Faster RCNN	ResNet50-FPN	44.3	46.8	73.7	36.0	55.8	26.9	37.8	58.0	52.7	48.0	+21.6

Table 3: Detection results on *Day-to-Night* (*BDD100k daytime clear* → *BDD100k nighttime clear*). Average precision (%) is reported in the target domain. Light and Sign stand for Traffic Light and Traffic Sign respectively.

Method	FID↓	precision↑	recall↑	density↑	coverage↑
Cityscapes → FoggyCityscapes					
Source Only	55.74	0.045	**0.764**	0.011	0.012
CycleGAN	87.89	0.256	0.459	0.101	0.127
DCLGAN	70.81	0.372	0.194	0.172	0.192
CUT	38.35	0.529	0.235	0.298	0.218
OA-FSUI2IT	**33.02**	**0.684**	0.213	**0.501**	**0.436**
BDD100k Day → Night					
Source Only	110.29	0.145	**0.433**	0.040	0.019
CycleGAN	75.17	0.516	0.128	0.312	0.236
DCLGAN	84.99	0.552	0.051	0.315	0.207
CUT	86.18	0.456	0.416	0.306	0.302
OA-FSUI2IT	**33.21**	**0.557**	0.322	**0.422**	**0.362**
KITTI → Cityscapes					
Source Only	59.08	0.069	**0.455**	0.017	0.064
CycleGAN	91.10	0.598	0.081	0.530	0.579
DCLGAN	90.49	0.523	0.069	0.371	0.404
CUT	60.45	0.362	0.121	0.188	0.330
OA-FSUI2IT	**55.32**	**0.749**	0.016	**0.926**	**0.629**

Table 4: Quantitative comparisons for the quality of the generated images in the target domain.

Experimental Results

We conduct experiments and compare our OA-FSUI2IT framework with the state-of-the-art cross-domain detection methods: *DAFRCNN* (Chen et al. 2018), *SWDA* (Saito et al. 2019), *DDMRL* (Kim et al. 2019b), *CycleGAN* (Zhu et al. 2017), *CUT* (Park et al. 2020), *DCLGAN* (Han et al. 2021) and *OSHOT* (Innocente et al. 2020). Quantitative results of OA-FSUI2IT for the translated images in comparison with CycleGAN, CUT, and DCLGAN are presented in Tab. 4.

Normal-to-Foggy: We achieve the state-of-the-art mAP of 42.3 in this FSCD detection task, in Tab. 1, with remarkable performance gains of 41.0% and 7.1% in comparison to Source Only and CUT baseline. We find OA-FSUI2IT network can guide the distribution of generated samples much closer to *FoggyCityscapes*, according to the qualitative and quantitative (reach lowest *FID*: 33.02) comparison results of the generated images presented in Fig. 5, and Tab. 4. Even in the One-Shot scenario, our OA-FSUI2IT surpass the strong baselines of OSHOT and CUT, as shown in Tab. 1.

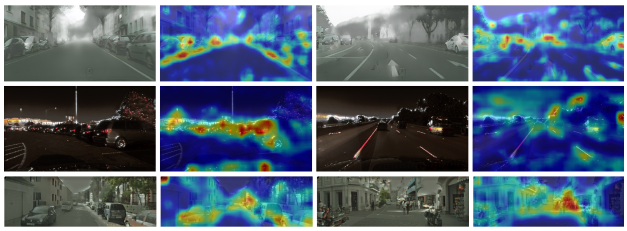

Figure 9: Visualization of domain evidence using Grad-CAM. First row: *Normal-to-Foggy*. Second row: *Day-to-Night*. Third row: *Cross scene adaptation*.

Day-to-Night: The Tab. 3 shows the FSCD detection results for the *Day-to-Night* task. We achieve an mAP of 30.5, with 4.1 and 4.2 higher than Source Only (26.4) and CUT (26.3) baselines. Much lower *FID* (33.21) than Source Only (110.29) and CUT (86.18), and qualitative comparisons of the generated images, presented in Fig. 6, also demonstrate the superior performance for our OA-FSUI2IT network.

Cross Scene Adaptation: We achieve the mAP of 26.3, with performance gains of 29.6% and 21.2% in comparison to the Source Only (20.3) and CUT (21.7) baseline, in Tab. 2. We procure lower FID (55.32) than Source Only (59.08) and CUT (60.45) indicating the domain disparity caused by the scenes variation has been narrowed, as shown in Fig. 7.

Further Empirical Analysis

Ablation Study. We implement the ablation study to investigate the effectiveness of each module in OA-FSUI2IT. We use the *Normal-to-Foggy* task as a study case, and the results are reported in Tab. 5. It shows the fact that when any one of the proposed modules is removed, the performance will drop correspondingly, demonstrating all the modules are designed reasonably. It is noteworthy that the degeneration of using self-supervised cycle-consistency module alone is predictable, as only implementing augmentation in the generator is harmful and may cause a performance drop. [3]

[3]Implementing data augmentation in unconditional GAN without involving discriminator lead to degradation (Karras et al. 2020).

Method	DA	PPCL	SSCC	person	rider	car	truck	bus	train	mcycle	bicycle	mAP	gain
CUT	✗	✗	✗	44.5	53.4	58.3	22.8	36.6	14.9	34.4	51.3	39.5	-
OA-FSUI2IT	✓	✗	✗	47.2	52.3	61.7	28.7	44.6	15.1	30.5	48.9	41.1	+1.6
OA-FSUI2IT	✗	**6**	✗	46.6	52.6	63.7	24.5	40.8	17.7	32.0	49.7	41.0	+1.5
OA-FSUI2IT	✗	✗	✓	47.5	53.8	59.0	21.9	39.4	17.9	27.1	49.0	39.0	-0.5
OA-FSUI2IT	✓	**6**	✗	47.0	54.6	64.6	25.9	40.5	15.3	31.6	51.3	41.3	+1.8
OA-FSUI2IT	✓	**6**	✓	46.5	51.8	62.6	28.1	40.9	20.9	30.8	49.3	41.4	+1.9
OA-FSUI2IT	✗	**9**	✗	44.9	51.3	63.0	27.7	46.1	18.2	34.3	48.2	41.7	+2.2
OA-FSUI2IT	✓	**9**	✓	47.5	53.8	64.1	27.8	45.9	11.5	35.9	52.3	42.3	+2.8

Table 5: Ablation study of the OA-FSUI2IT based FSCD object detection framework on *Normal-to-Foggy* (*Cityscapes* → *FoggyCityscapes*) using Faster RCNN with ResNet50-FPN backbone. Average precision (%) is reported in the target domain.

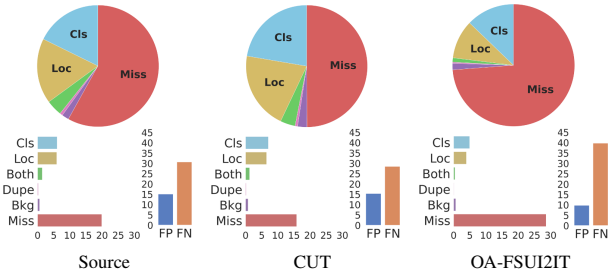

Figure 10: Error analysis on *Normal-to-Foggy* adaptation.

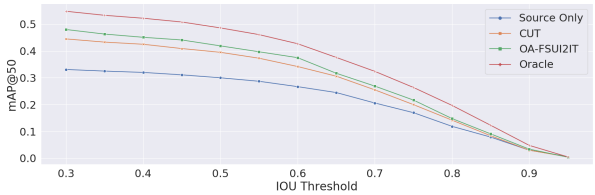

Figure 11: The performance with the variation of IOU thresholds on *Normal-to-Foggy* adaptation.

Influence of IOU Threshold. Fig. 11 shows the performance of different models with the variation of IOU thresholds. The mAP constantly drops with the increase of IOU threshold and approaches zero in the end. OA-FSUI2IT consistently outperforms the CUT baseline within the IOU range $[0.3, 0.95]$, illustrating that our OA-FSUI2IT provides more accurate and robust bounding boxes regression.

Detection Error Analysis. We employ TIDE (Bolya et al. 2020) to analyze the errors of Source Only, CUT, and OA-FSUI2IT. The detection errors are categorized into 6 main error types in TIDE: 1) Classification (*Cls*), 2) Localization (*Loc*), 3) both Cls and Loc (*Both*), 4) Duplicate detection (*Dup*), 5) Background (*Bkg*), and 6) Missed GT (*Miss*); and two separate error types: 1) False Positive (FP), and 2) False Negative (FN). The highest *confident* detections and lowest *Miss* detections on the *Normal-to-Foggy* task further verify the effectiveness of the proposed OA-FSUI2IT framework for FSCD object detection. We report the absolute and relative error contribution for each error type across all categories in Fig. 10. Comparing to Source Only and CUT, our OA-FSUI2IT network can generate more realistic fog, from which the *de facto* detectors can learn to distinguish the objects in the fog better. Thus, the OA-FSUI2IT framework clearly reduces the number of incorrect detections and false negatives. Meanwhile, as some objects may be occluded by the generated fog, the number of false positives increases slightly which could be thought as a reasonable sacrifice.

Qualitative Comparison of the Detection Results. Fig. 8 presents the qualitative comparison of detection results on transfer tasks, *Normal-to-Foggy*, *Day-to-Night*, and *Cross scene adaptation*. Our OA-FSUI2IT consistently outperforms Source Only and CUT. For example, in the first row of Fig. 8, both CUT and OA-FSUI2IT can successfully detect the cars occluded by the fog; meanwhile OA-FSUI2IT is able to locate more cars under the same circumstance. Also, the second row indicates that OA-FSUI2IT can identify the car in the dark, on the left side of the image, and the person beside the sliver SUV. The OA-FSUI2IT framework produces fewer false positives of traffic signs and traffic lights than CUT. In the last row, our method can detect more people on the sidewalk.

Visualization of Domain Evidence. In Fig. 9, we also use Grad-CAM (Selvaraju et al. 2017) to show the evidence (e.g., heatmap) to demonstrate the images are successfully transferred from source domains to target domains, for the adaptation of *Normal-to-Foggy*, *Day-to-Night*, and *Cross Scene*. The heatmaps show that our method provides reasonable focus on category features (i.e., car and person), which is beneficial to the translation results and the detector.

Conclusion

In this paper, we present a novel OA-FSUI2IT framework for FSCD object detection. To our knowledge, it is the first work to successfully address FSCD detection. Our key contributions include: 1) We propose the OA-FSUI2IT framework to address the FSCD object detection; 2) We present a series of new modules, i.e., discriminator augmentation, patch pyramid contrastive learning, and self-supervised content-consistency, to improve FSCD detection performance; 3) We perform extensive experiments and achieve state-of-the-art performance on multiple FSCD object detection benchmarks (e.g., *Cityscapes* → *FoggyCityscapes*, *BDD100k Daytime Clear*→ *Nighttime Clear*, *Kitti* → *Cityscapes*) to demonstrate its superiority.

Acknowledgments

This work was sponsored by the Shanghai Pujiang Program (20PJ1419000), Shanghai Rising-Star Program (20QB1405500), Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (2020ZDSYSKFKT03-1), Development and Application of Intelligent Diagnosis System for Mixed Traffic Characteristics (2019-RGZN-01023), and Development and Industrialization of Deep Neural Network based Intelligent Inspection System for Vehicle Safety and Compliance (190248).

References

Arruda, V. F.; Paixão, T. M.; Berriel, R. F.; De Souza, A. F.; Badue, C.; Sebe, N.; and Oliveira-Santos, T. 2019. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *IJCNN*, 1–8.

Bolya, D.; Foley, S.; Hays, J.; and Hoffman, J. 2020. TIDE: A general toolbox for identifying object detection errors. In *ECCV*, 558–573.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020a. Harmonizing Transferability and Discriminability for Adapting Object Detectors. In *CVPR*, 8869–8878.

Chen, R.; Huang, W.; Huang, B.; Sun, F.; and Fang, B. 2020b. Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Translation. In *CVPR*, 8168–8177.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 3339–3348.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

Doersch, C.; and Zisserman, A. 2017. Multi-task Self-Supervised Visual Learning. *ICCV*, 2070–2079.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, 1180–1189.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The kitti vision benchmark suite. In *CVPR*, 3354–3361.

Gopalan, R.; Li, R.; and Chellappa, R. 2011. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 999–1006.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 297–304.

Han, J.; Shoeiby, M.; Petersson, L.; and Armin, M. A. 2021. Dual Contrastive Learning for Unsupervised Image-to-Image Translation. In *CVPR workshop*, 746–755.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 6629—6640.

Innocente, A.; Borlino, F. C.; Bucci, S.; Caputo, B.; and Tommasi, T. 2020. One-Shot Unsupervised Cross-Domain Detection. In *ECCV*, 732–748.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 5967–5976.

Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2020. Neural Style Transfer: A Review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11): 3365–3385.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training Generative Adversarial Networks with Limited Data. In *NeurIPS*, 12104–12114.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019a. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019b. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 12456–12465.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*, 1–15.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*, 121–135.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 84–90.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NeurIPS*, 700–708.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-Shot Unsupervised Image-to-Image Translation. In *ICCV*, 10550–10559.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*, 21–37.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning Transferable Features with Deep Adaptation Networks. In *ICML*, 97–105.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *CVPR*, 2794–2802.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In *ICML*, 7176–7185.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pang, Y.; Lin, J.; Qin, T.; and Chen, Z. 2021. Image-to-Image Translation: Methods and Applications. *arXiv preprint arXiv:2101.08629*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *ECCV*, 319–345.

Patel, V. M.; Gopalan, R.; Li, R.; and Chellappa, R. 2015. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3): 53–69.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *CVPR*, 7263–7271.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

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

Saito, K.; Saenko, K.; and Liu, M.-Y. 2020. COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder. In *ECCV*, 382–398.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 6956–6965.

Sajjadi, M. S.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing generative models via precision and recall. In *NeurIPS*, 5234–5243.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9): 973–992.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

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

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *ECCV*, 443–450.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 7167–7176.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; kavukcuoglu, k.; Vinyals, O.; and Graves, A. 2016. Conditional Image Generation with PixelCNN Decoders. In *NeurIPS*, 4797–4805.

Wang, T.; Zhang, X.; Yuan, L.; and Feng, J. 2019. Few-Shot Adaptive Faster R-CNN. *CVPR*, 7166–7175.

Wilson, G.; and Cook, D. J. 2020. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5): 1–46.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR*, 2636–2645.

Yu, F.; Wang, D.; Chen, Y.; Karianakis, N.; Shen, T.; Yu, P.; Lymberopoulos, D.; Lu, S.; Shi, W.; and Chen, X. 2019. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*.

Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; and Saminger-Platz, S. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*.

Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable Augmentation for Data-Efficient GAN Training. In *NeurIPS*, 7559–7570.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2223–2232.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.