International Journal of Computer Vision (2021) 129:761–780
https://doi.org/10.1007/s11263-020-01394-z

CDTD: A Large-Scale Cross-Domain Benchmark for Instance-Level Image-to-Image Translation and Domain Adaptive Object Detection

Zhiqiang Shen[1] · Mingyang Huang[2] · Jianping Shi[2] · Zechun Liu[1] · Harsh Maheshwari[1] · Yutong Zheng[1] · Xiangyang Xue[3] · Marios Savvides[1] · Thomas S. Huang[4]

Received: 15 March 2020 / Accepted: 14 October 2020 / Published online: 24 November 2020

Abstract
Cross-domain visual problems, such as image-to-image translation and domain adaptive object detection, have attracted increasing attentions in the last few years, and also become new rising and challenging directions for the computer vision community. Recently, despite enormous efforts of the field in data collection, there are still few datasets covering the instance-level image-to-image translation and domain adaptive object detection tasks simultaneously. In this work, we introduce a large-scale cross-domain benchmark **CDTD** (contains 155,529 high-resolution natural images across four different modalities with object bounding box annotations. A summary of the entire dataset is provided in the following sections. Dataset is available at: http://zhiqiangshen.com/projects/INIT/index.html.) for the new instance-level translation and object detection tasks. We provide comprehensive baseline results of the benchmark on both of these two tasks. Moreover, we proposed a novel instance-level image-to-image translation approach called INIT and a gradient detach method for the domain adaptive object detection to harvest and exert dataset's function of the instance level annotations across different domains.

Keywords Cross-domain benchmark · Instance level image-to-image translation · Domain adaptive object detection

1 Introduction

In real world scenarios, generic vision tasks like image recognition, object detection, image translation, etc., always face severe challenges from variations in viewpoint, background, object appearance, illumination, occlusion conditions, scene change, etc. These unavoidable factors make these tasks in domain-shift circumstance a challenging and new rising research topic in the recent years. Also, domain change is a widely-recognized, intractable problem that urgently needs to break through in reality tasks, like video surveillance, autonomous driving, etc. Consequently, a large-scale cross-domain benchmark is urgently-needed for pushing this field forward.

The recent emergence of large-scale image datasets in the cross-domain circumstance like VisDA (Peng et al. 2017), Office-Home (Venkateswara et al. 2017), Syn2real (Peng et al. 2018), DomainNet (Peng et al. 2019) are mainly focusing on the traditional classification or detection tasks, thus they are not flexible to be applied to new raised tasks

Communicated by Dengxin Dai.

✉ Zhiqiang Shen
zhiqians@andrew.cmu.edu

Mingyang Huang
huangmingyang@sensetime.com

Jianping Shi
shijianping@sensetime.com

Zechun Liu
zechunl@andrew.cmu.edu

Yutong Zheng
yutongzh@andrew.cmu.edu

Xiangyang Xue
xyxue@fudan.edu.cn

Marios Savvides
marioss@andrew.cmu.edu

Thomas S. Huang
t-huang1@illinois.edu

[1] Carnegie Mellon University, Pittsburgh, USA

[2] SenseTime Research, Beijing, China

[3] Fudan University, Shanghai, China

[4] University of Illinois at Urbana-Champaign, Champaign, USA

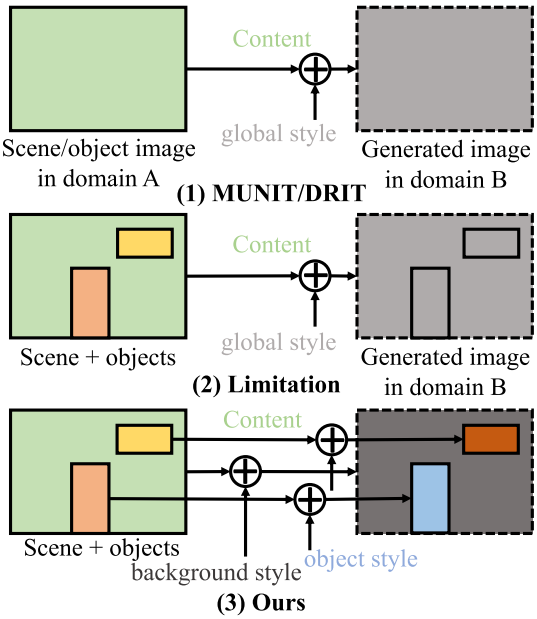

Fig. 1 Illustration of the motivation of our method. (1) MUNIT (Huang et al. 2018)/DRIT (Lee et al. 2018) methods; (2) their limitation; and (3) our solution for instance-level translation. More details can be referred to the text

like image-to-image translation, especially the instance level translation task. The motivation of this work is to build a dataset that has instance-level annotations of images (every instance has a bounding box coordinate and a semantic label) under a large, unrestricted and real world scenarios across different domains, in order to solve the instance-level image-to-image translation and further extend to domain adaptive object detection tasks.

Instance-level image-to-image translation Image-to-Image (I2I) translation has become more and more important in computer vision recently, since many vision and graphics problems can be formulated as an I2I translation problem like super-resolution, neural style transfer, colorization, etc. This technique has also been adapted to the relevant fields such as medical image processing (Zhang et al. 2018) to further improve the medical volumes segmentation performance. In general, Pix2pix (Isola et al. 2017) is regarded as the first unified framework for I2I translation which adopts conditional generative adversarial networks (Mirza and Osindero 2014) for image generation, while it requires the paired examples during training process. A more general and challenging setting is the unpaired I2I translation, where the paired data is unavailable.

Several recent efforts (Zhu et al. 2017; Liu et al. 2017; Huang et al. 2018; Lee et al. 2018; Almahairi et al. 2018) have been made on this direction and achieved very promising results. For instance, CycleGAN (Zhu et al. 2017) proposed the cycle consistency loss to enforce the learning process that if an image is translated to the target domain by learning a mapping and translated back with an inverse mapping, the output should be the original image. Furthermore, CycleGAN assumes the latent spaces are separate from the two mappings. In contrast, UNIT (Liu et al. 2017) assumes two domain images can be mapped onto a shared latent space. MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018) further postulate that the latent spaces can be disentangled to a shared content space and a domain-specific attribute space.

However, all of these methods thus far have focused on migrating styles or attributes onto the entire images. As shown in Fig. 1 (1), they work well on the unified-style scenes or relatively content-simple scenarios due to the consistent pattern across various spatial areas in an image, while this is not true for the complex structure images with multiple objects since the stylistic vision disparity between objects and background in an image is always huge or even totally different, as in Fig. 1 (2).

To address the aforementioned limitation, in this work we present a method that can translate objects and background/global areas separately with different style codes as in Fig. 1 (3), and still training in an end-to-end manner. The motivation of our method is illustrated in Fig. 3. Instead of using the global style, we use instance-level style vectors that can provide more accurate guidance for visually related object generation in target domain. We argue that styles should be diverse for different objects, backgrounds or global images, meaning that the style codes should not be identical for the entire image. More specifically, a car from "sunny" to the "night" domain should have different style codes comparing to the global image translation between these two domains. Our method achieves this goal by involving the instance-level styles. Given a pair of unaligned images and object locations, we first apply our encoders to obtain the intermediate global and instance level content and style vectors separately. Then we utilize the cross-domain mapping to obtain the target domain images by swapping the style/attribute vectors. Our swapping strategy is introduced with more details in Sect. 4. The main advantage of our method is the exploration and usage of object level styles, which affects and guides the generation of target domain objects directly. Certainly, we can also apply the global style for target objects to enforce the model to learn more diverse results.

Domain adaptive object detection As illustrated in Fig. 2, unsupervised domain adaptive object detection aims to learn a robust detector in the domain shift circumstance, where the training (source) domain is label-rich with bounding box annotations, while the testing (target) domain is label-agnostic and the feature distributions between training and testing domains are dissimilar or even totally different. Previous solutions on this problem usually design distribution alignments on global and local level images by using an adversarial loss. The alignments generally require additional

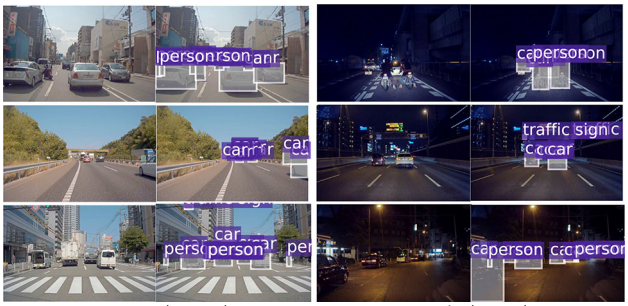

Fig. 2 Illustration of domain-shift object detection in autonomous driving scenario. Images are from our CDTD dataset (Shen et al. 2019)

components or sub-networks to realize, which are troublesomely complicated and poorly interpretable. In this work, we propose a simple training technique called *gradient detach* that prevents the flow of gradients from context sub-network through the detection backbone path, so that it can learn more discriminative representations between object and global/context images, and focus more on the target areas. After accompanying with the compatible stacked complementary losses by cutting in several auxiliary objectives in different network stages, our method can automatically align the distributions of source and target domains effectively. We conduct experiments on the proposed dataset with two baseline methods DA (Chen et al. 2018) and strong-weak alignment (Saito et al. 2019), our results are consistently better than the two baseline methods.

In summary, our contributions are four fold:

- We introduce a large-scale, multimodal, highly varied and high-resolution cross domain dataset, containing ∼155k streetscape images across four domains. Our dataset not only includes the domain category labels, but also provides the detailed object bounding box annotations, which will benefit the instance-level I2I translation and domain adaptive object detection problems.
- We propel I2I translation problem step forward to instance-level such that the constraints could be exploited on both instance and global-level attributes by adopting the proposed compound loss.
- We conduct extensive qualitative and quantitative experiments to demonstrate that our approach can surpass the baseline I2I translation methods. Our synthetic images can be even beneficial to other vision tasks such as generic object detection, and further improve the performance.
- We propose a novel training strategy, *gradient detach*, for the domain adaptive object detection task which suppresses gradients flowing back to the detection backbone. To our best knowledge, this may be the first work to show the effectiveness of gradient detach that can help to learn better context representation for domain adaptive object detection. In addition, we proposed to use multiple complementary losses to help gradient detach training for better optimization.

A preliminary version (Shen et al. 2019) of this manuscript has been published in a previous conference CVPR 2019. Compared to the previous conference paper, our major new contributions are that we extend our dataset to domain adaptive object detection task, we propose a gradient detach based stacked complementary losses approach to boost the previous state-of-the-art methods and achieve fairly competitive performance. We also conduct additional experiments and visualizations on the original instance-level image-to-image translation task. Moreover, we include more description of the dataset, the method for domain adaptive object detection and more baseline results.

The rest sections of this work are organized as follows. In Sect. 2, we review the related work of our study. In Sect. 3, we introduce the construction of the CDTD dataset and its statistics. We also provide a feature-by-feature comparison to other related datasets. In Sect. 4, we introduce the proposed INIT method for instance-level image-to-image translation. We propose to use the fine-grained local (instance) and global styles on the target image spatially to translate the source images. In Sect. 5, we introduce a gradient detach method for the domain adaptive object detection task. The proposed method prevents the flow of gradients from context sub-network through the detection backbone path, so that it can learn more discriminative representations between object and global/context images, and focus more on the target areas. In Sect. 6 we provide extensive experiments and ablation studies on our collected dataset of image-to-image translation task, some baselines and our method results on domain adaptive object detection task. Sect. 7 concludes this work.

2 Related Work

2.1 Cross Domain Datasets for Translation and Object Detection

A variety of datasets have been collected for the purpose of cross domain study. In image-to-image translation field, the most commonly used ones are edge ↔ shoes (Isola et al. 2017), Yosemite (summer ↔ winter) (Zhu et al. 2017), Cityscapes (Cordts et al. 2016), while as shown in Table 1, these datasets either are in low-resolution (e.g., edge ↔ shoes), or have limited scale, i.e., number of images is too small (e.g., Cityscapes). In contrast, our dataset has more sufficient images to explore the potential of proposed algorithm. As shown in Table 2, the central weakness of current domain adaptive object detection datasets is the scale in terms of the number of images. In general, our dataset is about 15∼20×

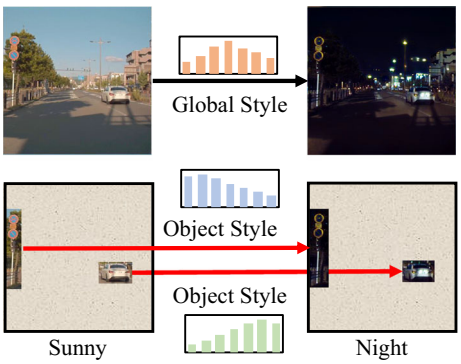

Fig. 3 A natural image example of our I2I translation

larger than these existing ones with higher quality/resolution of images.

Image-to-Image Translation The goal of I2I translation is to learn the mapping between two different domains. Pix2pix (Isola et al. 2017) first proposes to use conditional generative adversarial networks (Mirza and Osindero 2014) to model the mapping function from input to output images. Inspired by Pix2pix, some works further adapt it to a variety of relevant tasks, such as semantic layouts → scenes (Karacan et al. 2016), sketches → photographs (Sangkloy et al. 2017), etc. Despite popular usage, the major weaknesses of these methods are that they require the paired training examples and the outputs are single-modal. In order to produce multimodal and more diverse images, BicycleGAN (Zhu et al. 2017) encourages the bijective consistency between the latent and target spaces to avoid the mode collapse problem. A generator learns to map the given source image, combined with a low-dimensional latent code, to the output during training. While this method still needs the paired training data.

Recently, CycleGAN (Zhu et al. 2017) is proposed to tackle the unpaired I2I translation problem by using the cycle consistency loss. UNIT (Liu et al. 2017) further makes a share-latent assumption and adopts Coupled GAN in their method. To address multimodal problem, MUNIT (Huang et al. 2018), DRIT (Lee et al. 2018), Augmented CycleGAN (Almahairi et al. 2018), etc. adopt a disentangled representation to further learn diverse I2I translation from unpaired training data.

Instance-level Image-to-Image Translation To the best of our knowledge, there are so far very few efforts on the instance-level I2I translation problem. Perhaps the most similar to our work is the recently proposed *Insta*GAN (Mo et al. 2019), which utilizes the object segmentation masks to translate both an image and the corresponding set of instance attributes while maintaining the permutation invariance property of instances. A context preserving loss is designed to encourage model to learn the identity function outside of target instances. The main difference with ours is that *insta*GAN cannot translate different domains for an entire image sufficiently. They focus on translating instances and maintain the outside areas, in contrast, our method can translate instances and outside areas simultaneously and make global images more realistic. Furthermore, *Insta*GAN is built on the CycleGAN (Zhu et al. 2017), which is single modal, while we choose to leverage the MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018) to build our INIT, thus our method inherits multimodal and unsupervised properties, meanwhile, produces more diverse and higher quality images.

Some other existing works (Ma et al. 2018; Li et al. 2018) are more or less related to this paper. For instance, DA-GAN (Ma et al. 2018) learns a deep attention encoder to enable the instance-level translation, which is unable to handle the multi-instance and complex circumstance. BeautyGAN (Li et al. 2018) focuses on facial makeup transfer by employing histogram loss with face parsing mask. Mechrez et al. Mechrez et al. (2018) proposed a contextual loss based on the images' context and semantics, which compared regions with similar semantic information, meanwhile, considering the context of the entire image.

Domain Adaptive Object Detection. Unsupervised domain adaptation for recognition has been widely studied by a large body of previous literature (Ganin et al. 2016; Long et al. 2016; Tzeng et al. 2017; Panareda Busto and Gall 2017; Hoffman et al. 2018; Murez et al. 2018; Zhao et al. 2019; Wu et al. 2019), our method more or less draws merits from them, like aligning source and target distributions with adversarial learning (domain-invariant alignment). However, object detection is a technically different problem from classification, since we would like to focus more on the object of interests (regions).

Common approaches for tackling domain-shift object detection are mainly in two directions: (i) training supervised model and then fine-tuning on the target domain; or (ii) unsupervised cross-domain representation learning. The former requires additional instance-level annotations on target data, which is fairly laborious, expensive and time-consuming. So most approaches focus on the latter one but still have some challenges. The first challenge is that the representations of source and target domain data should be embedded into a common space for matching the object, such as the hidden feature space (Saito et al. 2019; Chen et al. 2018), input space (Tzeng et al. 2018; Cai et al. 2019) or both of them (Kim et al. 2019). The second is that a feature alignment or matching operation or mechanism for source/target domains should be further defined, such as subspace alignment (Raj et al. 2015), $\mathcal{H}$-divergence and adversarial learning (Chen et al. 2018), MRL (Kim et al. 2019), Strong-Weak alignment (Saito et al. 2019), universal alignment (Wang et al. 2019), etc. In general, our proposed method in this work targets at these two challenges, and it is also a learning-based alignment method across domains with an end-to-end framework.

Fig. 4 Image samples from our benchmark grouped by their domain categories (sunny, night, cloudy and rainy). In each group, left are original images and right are images with corresponding bounding box annotations

Table 1 Feature-by-feature comparison of popular I2I translation datasets

Datasets	Paired	Resolution	Bbox annotations	Modalities	# images
edge↔shoes (Isola et al. 2017)	✓	Low	–	{edge, shoes}	50,000
edge↔handbags (Isola et al. 2017)	✓	Low	–	{edge, handbags}	137,000
CMP Facades (Radim Tyleček 2013)	✓	HD	–	{facade, semantic map}	606
Yosemite (summer↔winter) (Zhu et al. 2017)	✗	HD	–	{summer, winter}	2127
Yosemite* (MUNIT) (Huang et al. 2018)	✗	HD	–	{summer, winter}	5638
Cityscapes (Cordts et al. 2016)	✓	HD	✓	{ semantic, realistic}	3475
Transient Attributes (Laffont et al. 2014)	✓	HD	✗	{40 transient attributes}	8571
Ours	✗	HD†	✓	{sunny, night, cloudy, rainy}	155,529

Our dataset contains four relevant but visually-different domains: sunny, night, cloudy and rainy.
†indicates that the images in our dataset contain two types of resolutions: 1208×1920 and 3000×4000

Table 2 Comparison of popular domain adaptive object detection datasets

Datasets	Resolution	Modalities	Train/test in source	Train/test in target
Cityscapes (Cordts et al. 2016)→FoggyCityscapes (Sakaridis et al. 2018)	HD	{normal, foggy}	2975/500	2975/500
Cityscapes (Cordts et al. 2016)→KITTI (Geiger et al. 2012)	HD	{real, real}	2975/500	7481/7518
KITTI (Geiger et al. 2012)→Cityscapes (Cordts et al. 2016)	HD	{real, real}	7481/7518	2975/500
PASCAL* (Everingham et al. 2010)→Clipart1k (Inoue et al. 2018)	HD	{real, cartoon}	16,551/15,943	1000/–
PASCAL* (Everingham et al. 2010)→WaterColor2k (Inoue et al. 2018)	HD	{real, artistic}	16,551/15,943	2000/–
GTA (Sim10K) (Johnson-Roberson et al. 2016)→Cityscapes (Cordts et al. 2016)	HD	{ synthetic, real}	10,000/–	2975/500
Ours	HD†	{sunny, night, cloudy, rainy}	Total: 155,529	

*denotes the `2007+2012 trainval` combination of PASCAL VOC dataset.
†indicates that the images in our dataset contain two types of resolutions: 1208×1920 and 3000×4000

Table 3 Statistics (# images) of the entire dataset across four domains: sunny, night, rainy and cloudy

Domain	Training (85%)	Testing (15%)	Total (100%)
Sunny	49,663	8764	58,427
Night	24,559	4333	28,892
Rainy	6041	1066	7107
Cloudy	51,938	9165	61,103
Total	132,201	23,328	155,529

The data is divided into two subsets: 85% for training and 15% for testing

3 CDTD: A Cross-Domain Dataset with Instance Bounding-box Annotations

We introduce a large-scale street scene centric dataset CDTD[1] that addresses three core research problems in I2I translation: (1) unsupervised learning paradigm, meaning that there is no specific one-to-one mapping in the data; (2) multimodal domains incorporation. Most existing I2I translation datasets provide only two different domains, which limit the potential to explore more challenging tasks like multi-domain incorporation circumstance. Our dataset contains four domains: sunny, night, cloudy and rainy[2] in a unified street scene; and (3) multi-granularity (global and instance-level) information. Our dataset provides instance-level bounding box annotations, which can utilize more details for learning a translation model. Table 1 shows a feature-by-feature comparison among various I2I translation datasets. We also visualize some examples of the dataset in Fig. 4. For instance category, we annotate three common objects in street scenes including: car, person, traffic sign (speed limited sign). As our dataset covers multiple domains with shared categories, so it is also suitable for the domain adaptive object detection task.

3.1 Dataset summary

CDTD dataset consists of 155,529 images, among it, there are 132,201 images for training and 23,328 images for testing. The dataset contains four relevant but visually-different domains: sunny, night, cloudy, rainy. The detailed statistics (#images) of the entire dataset are shown in Table 3. All the images are collected in Tokyo, Japan with SEKONIX AR0231 camera. The whole collection process lasted about 3 months.

[1] The abbreviation of *A* **C***ross*-**D***omain Benchmark for* **T***ranslation and* **D***etection tasks.*

[2] For safety, we collect the rainy images after the rain, so this category looks more like overcast weather with wet road.

4 Instance-aware Image-to-Image Translation

Unpaired Image-to-image Translation aims to learn a mapping between unaligned image pairs in diverse domains. Recent advances in this field like MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018) mainly focus on disentangling content and style/attribute from a given image first, then directly adopting the global style to guide the model to synthesize new domain images. However, this kind of approaches severely incurs contradiction if the target domain images are content-rich with multiple discrepant objects. In this paper, we present a simple yet effective instance-aware image-to-image translation approach (INIT), which employs the fine-grained local (instance) and global styles to the target image spatially. The proposed INIT exhibits three import advantages: (1) the instance-level objective loss can help learn a more accurate reconstruction and incorporate diverse attributes of objects; (2) the styles used for target domain of local/global areas are from corresponding spatial regions in source domain, which intuitively is a more reasonable mapping; (3) the joint training process can benefit both fine and coarse granularity and incorporates instance information to improve the quality of global translation. We observe that our synthetic images can even benefit real-world vision tasks like generic object detection.

More precisely, our goal is to realize the instance-aware I2I translation between two different domains without paired training examples. We build our framework by leveraging the MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018) methods. To avoid repetition, we omit some innocuous details. Similar to MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018), our method is straight-forward and simple to implement. As illustrated in Fig. 6, our translation model consists of two encoders E_g, E_o (g and o denote the global and instance image regions respectively), and two decoders G_g, G_o in each domain $\mathcal{X}$ or $\mathcal{Y}$. Since we have the object coordinates, we can crop the object areas and feed them into the instance-level encoder to extra the content/style vectors. An alternative method for object content vectors is to adopt RoI pooling (Girshick 2015) from the global image content features. Here we use image crop (object region) and share

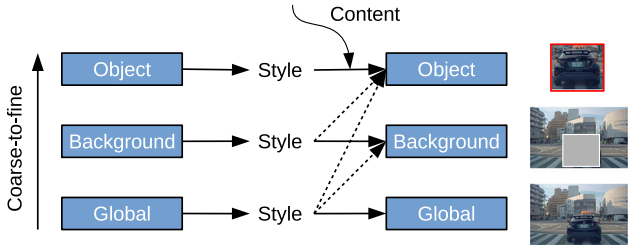

Fig. 5 Our content-style pair association strategy. Only coarse styles can be applied to fine contents, the reversal of processing flow is not allowed during training

the parameters for the two encoders, which is easier to implement.

Disentangle content and style on object and entire image. As (Cheung et al. 2015; Mathieu et al. 2016; Huang et al. 2018; Lee et al. 2018), our method also decomposes input images/objects into a shared content space and a domain-specific style space. Take global image as an example, each encode E_g can decompose the input to a content code c_g and a style code s_g, where $E_g = (E_g^c, E_g^s)$, $c_g = E_g^c(I)$, $s_g = E_g^s(I)$, I denotes the input image representation. c_g and s_g are global-level content/style features.

Generate style code bank. We generate the style codes from objects, background and entire images, which form our style code bank for the following swapping operation and translation. In contrast, MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018) use only the entire image style or attribute, which is struggling to model and cover the rich image spatial representation.

Associate content-style pairs for cyclic reconstruction. Our cross-cycle consistency is performed by swapping encoder-decoder pairs (dashed arc lines in Fig. 7). The cross-cycle includes two modes: cross-domain ($\mathcal{X} \leftrightarrow \mathcal{Y}$) and cross-granularity (entire image $\leftrightarrow$ object). We illustrate cross-granularity (image $\leftrightarrow$ object) in Fig. 7, the cross-domain consistency ($\mathcal{X} \leftrightarrow \mathcal{Y}$) is similar to MUNIT (Huang et al. 2018) and DRIT (Lee et al. 2018). As shown in Fig. 5, the swapping or content-style association strategy is a hierarchical structure across multi-granularity areas. Intuitively, the coarse (global) style can affect fine content and be adopted to local areas, while it's not true if the process is reversed. Following (Huang et al. 2018), we also use AdaIN (Huang and Belongie 2017) to combine the content and style vectors, which can be formulated as:

$$\text{AdaIN}(c, s) = \sigma(s)\left(\frac{c - \mu(c)}{\sigma(c)}\right) + \mu(s) \tag{1}$$

where c is the input content batch, s is a style input. $\mu(c)$, $\sigma(c)$ are the mean and standard deviation and AdaIN aims to scale the normalized content input with $\sigma(s)$, and shift it with $\mu(s)$.

Incorporate Multi-Scale. It is technically easy to incorporate multi-scale advantage into the framework. We simply replace the object branch in Fig. 7 with resolution-reduced images. In our experiments, we use a 1/2 scale and original size images as pairs to perform scale-augmented training. Specifically, styles from the small size and original size images can be performed to each other, and the generator needs to learn multi-scale reconstruction for both of them, which leads to more accurate results.

Reconstruction loss. We use self-reconstruction and cross-cycle consistency loss (Lee et al. 2018) for both entire image and object that encourage reconstruction of them. With encoded c and s, the decoders should decode them back to original input,

$$\hat{I} = G_g(E_g^c(I),\ E_g^s(I)),\ \hat{o} = G_o(E_o^c(o),\ E_o^s(o)) \tag{2}$$

We can also reconstruct the latent distribution (i.e. content and style vectors) as (Huang et al. 2018).

$$\hat{c}_o = E_o^c(G_o(c_o, s_g)),\ \hat{s}_o = E_o^s(G_o(c_o, s_g)) \tag{3}$$

where c_o and s_g are instance-level content and global-level style features. Then, we can use the following formation to learn a reconstruction of them:

$$\mathcal{L}_{recon}^k = \mathbb{E}_{k\sim p(k)}\left[\left\|\hat{k} - k\right\|_1\right] \tag{4}$$

where k can be I, o, c or s. $p(k)$ denotes the distribution of data k. The formation of cross-cycle consistency is similar to this process and more details can be referred to (Lee et al. 2018).

Adversarial loss. Generative adversarial learning (Goodfellow et al. 2014) has been adapted to many visual tasks, e.g., detection (Nguyen et al. 2017; Bai et al. 2018), inpainting (Pathak et al. 2016; Iizuka et al. 2017; Yu et al. 2018), ensemble (Shen et al. 2019), etc. We adopt adversarial loss $\mathcal{L}_{adv}$ where $D_{\mathcal{X}}^g$, $D_{\mathcal{X}}^o$, $D_{\mathcal{Y}}^g$ and $D_{\mathcal{Y}}^o$ attempt to discriminate between real and synthetic images/objects in each domain. We explore two designs for the discriminators: weight-sharing or weight-independent for global and instance images in each domain. The ablation experimental results are shown in Tables 4 and 5, we observe that shared discriminator is a better choice in our experiments.

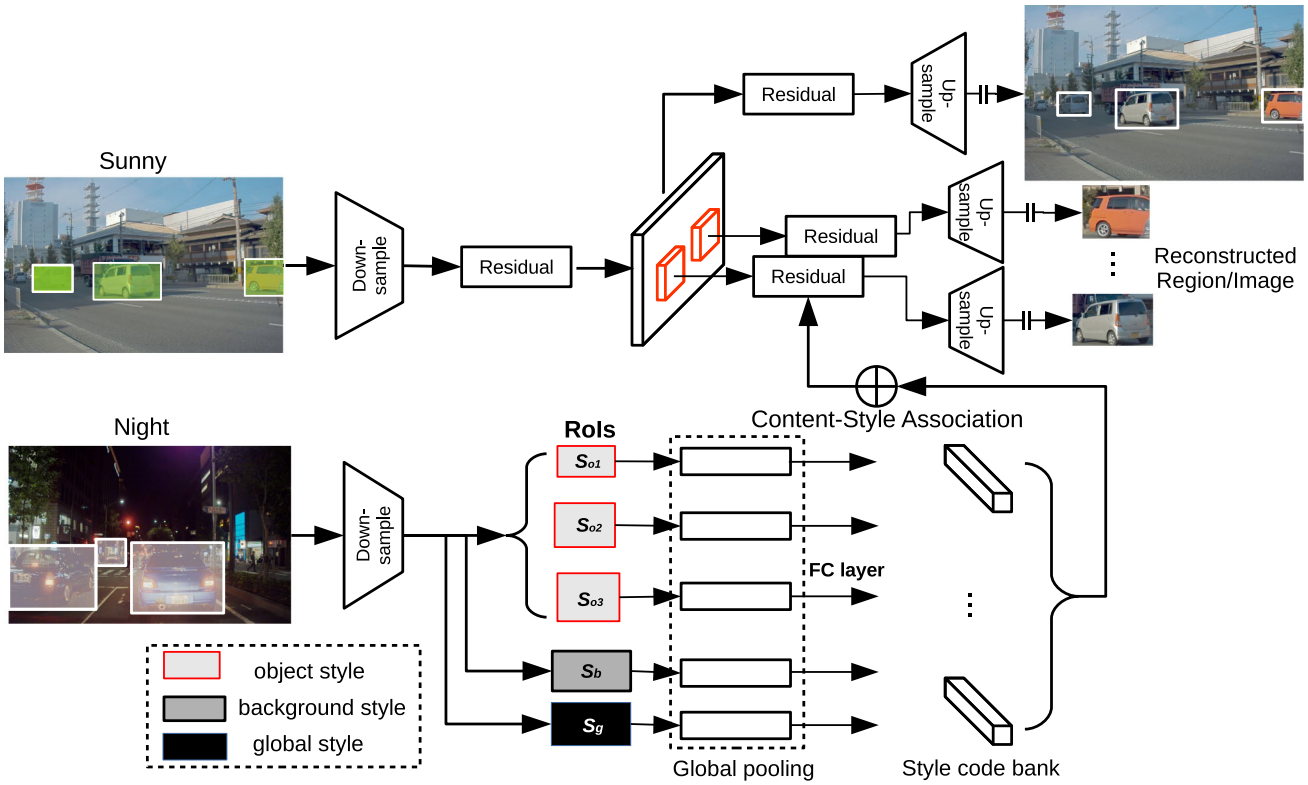

Fig. 6 Overview of our instance-aware cross-domain I2I translation. The whole framework is based on the MUNIT method (Huang et al. 2018), while we further extend it to realize the instance-level translation purpose. Note that after content-style association, the generated images will place in the target domain, so a translation back process will be employed before self-reconstruction, which is not illustrated here

Full objective function. The full objective function of our framework is:

$$\begin{aligned}
&\min_{E_{\mathcal{X}},E_{\mathcal{Y}},G_{\mathcal{X}},G_{\mathcal{Y}}} \max_{D_{\mathcal{X}},D_{\mathcal{Y}}} \mathcal{L}(E_{\mathcal{X}}, E_{\mathcal{Y}}, G_{\mathcal{X}}, G_{\mathcal{Y}}, D_{\mathcal{X}}, D_{\mathcal{Y}}) \\
&= \underbrace{\lambda_g(\mathcal{L}^{g\mathcal{X}} + \mathcal{L}^{g\mathcal{Y}}) + \lambda_{c_g}(\mathcal{L}_g^{c\mathcal{X}} + \mathcal{L}_g^{c\mathcal{Y}}) + \lambda_{s_g}(\mathcal{L}_g^{s\mathcal{X}} + \mathcal{L}_g^{s\mathcal{Y}})}_{global-level\ reconstruction\ loss} \\
&+ \underbrace{\lambda_o(\mathcal{L}^{o\mathcal{X}} + \mathcal{L}^{o\mathcal{Y}}) + \lambda_{c_o}(\mathcal{L}_o^{c\mathcal{X}} + \mathcal{L}_o^{c\mathcal{Y}}) + \lambda_{s_o}(\mathcal{L}_o^{s\mathcal{X}} + \mathcal{L}_o^{s\mathcal{Y}})}_{instance-level\ reconstruction\ loss} \\
&+ \underbrace{\mathcal{L}_{adv}^{\mathcal{X}_g} + \mathcal{L}_{adv}^{\mathcal{Y}_g}}_{global-level\ GAN\ loss} + \underbrace{\mathcal{L}_{adv}^{\mathcal{X}_o} + \mathcal{L}_{adv}^{\mathcal{Y}_o}}_{instance-level\ GAN\ loss}
\end{aligned} \tag{5}$$

where λ_g, λ_o, λ_{c_g}, λ_{c_p}, λ_{s_g}, λ_{s_o} are weights that control the importance of different reconstruction terms.

During inference time, we simply use the global branch to generate the target domain images (See Fig. 6 upper-right part) so that it is not necessary to use bounding box annotations at this stage, and this strategy can also guarantee that the generated images are harmonious.

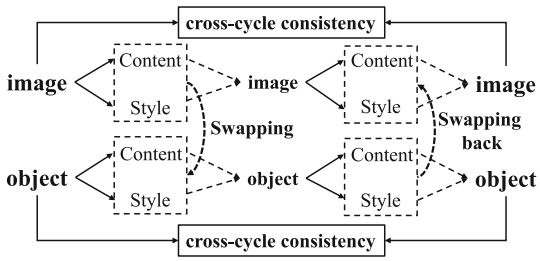

Fig. 7 Illustration of our cross-cycle consistency process. We only show cross-granularity (image ↔ object), the cross-domain consistency ($\mathcal{X} \leftrightarrow \mathcal{Y}$) is similar to the above paradigm

5 Domain Adaptive Object Detection

Unsupervised domain adaptive object detection aims to learn a robust detector in the domain shift circumstance, where the training (source) domain is label-rich with bounding box annotations, while the testing (target) domain is label-agnostic and the feature distributions between training and testing domains are dissimilar or even totally different.

Following the common formulation of domain adaptive object detection, we define a source domain $\mathcal{X}$ where annotated bounding-box is available, and a target domain $\mathcal{Y}$ where

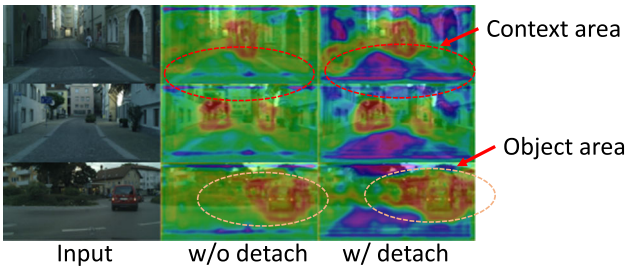

Fig. 8 Gradient detach helps to amplify contrast between context and object areas in domain adaptation scenario

only the image can be used in training process without any labels (bounding box and category). Our purpose is to train a robust detector that can adapt well to both source and target domain data, i.e., we aim to learn a *domain-invariant* feature representation that works well for detection across two different domains.

5.1 Gradient Detach Updating

In this section, we first introduce the detach strategy and how it helps to prevent the flow of gradients from context sub-network through the detection backbone path. Then we introduce the whole framework that we incorporate detach-based multi-objective learning on domain adaptive object detection scenario.

We define a sub-network to generate the context information from early layers of detection backbone. Intuitively, instance and context will focus on perceptually different parts of an image, so the representations from either of them should also be discrepant. However, if we train with the conventional joint process, the companion sub-network will be updated simultaneously with the detection backbone, which may lead to learning an indistinguishable representation/behavior from these two parts. To this end, in this work we propose to suppress gradients during backpropagation and force the representation of context sub-network to be dissimilar to the detection network, as shown in Algorithm 1. We then apply an instance-context alignment module with detach-generated context and backbone object representations for joint adaptation, as we elaborate in the following section. We find that gradient detach can help to obtain more discriminative context and object representations (see Fig. 8), and we show empirical evidence that this path carries information with diversity and hence gradients from this path getting suppressed is superior for such task.

Detach-Based Multi-Objective Learning. As shown in Fig. 9, we focus on the detach based complement objective learning and let $\mathcal{S} = \{(\mathbf{x}_i^{(\mathcal{X})}, \mathbf{y}_i^{(\mathcal{X})})\}$ where $\mathbf{x}_i^{(\mathcal{X})} \in \mathcal{R}^n$ denotes an image, $\mathbf{y}_i^{(\mathcal{X})}$ is the corresponding bounding box and category labels for sample $\mathbf{x}_i^{(\mathcal{X})}$, and i is an index. Each label $\mathbf{y}^{(\mathcal{X})} = (y_{\mathbf{c}}^{(\mathcal{X})}, y_{\mathbf{b}}^{(\mathcal{X})})$ denotes a class label $y_{\mathbf{c}}^{(\mathcal{X})}$ where $\mathbf{c}$ is the category, and a 4-dimension bounding-box coordinate $y_{\mathbf{b}}^{(\mathcal{X})} \in \mathcal{R}^4$. For the target domain we only use image data for training, so $\mathcal{T} = \{\mathbf{x}_i^{(\mathcal{Y})}\}$. We define a recursive function for layers $\mathbf{k} = 1, 2, \ldots, \mathbf{K}$ where we cut in complementary losses:

$$\hat{\Theta}_{\mathbf{k}} = \mathcal{F}\left(\mathbf{Z}_{\mathbf{k}}\right), \text{ and } \mathbf{Z}_0 \equiv \mathbf{x} \tag{6}$$

where $\hat{\Theta}_{\mathbf{k}}$ is the feature map produced at layer $\mathbf{k}$, $\mathcal{F}$ is the function to generate features at layer $\mathbf{k}$ and $\mathbf{Z}_{\mathbf{k}}$ is input at layer $\mathbf{k}$. We formulate the complement loss of domain classifier $\mathbf{k}$ as follows:

$$\begin{aligned}\mathcal{L}_{\mathbf{k}}\left(\hat{\Theta}_{\mathbf{k}}^{(\mathcal{X})}, \hat{\Theta}_{\mathbf{k}}^{(\mathcal{Y})}; \mathbf{D}_{\mathbf{k}}\right) &= \mathcal{L}_{\mathbf{k}}^{(\mathcal{X})}(\hat{\Theta}_{\mathbf{k}}^{(\mathcal{X})}; \mathbf{D}_{\mathbf{k}}) + \mathcal{L}_{\mathbf{k}}^{(\mathcal{Y})}(\hat{\Theta}_{\mathbf{k}}^{(\mathcal{Y})}; \mathbf{D}_{\mathbf{k}}) \\ &= \mathbb{E}\left[\log\left(\mathbf{D}_{\mathbf{k}}\left(\hat{\Theta}_{\mathbf{k}}^{(\mathcal{X})}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - \mathbf{D}_{\mathbf{k}}\left(\hat{\Theta}_{\mathbf{k}}^{(\mathcal{Y})}\right)\right)\right]\end{aligned} \tag{7}$$

where $\mathbf{D}_{\mathbf{k}}$ is the $\mathbf{k}$-th domain classifier or discriminator. $\hat{\Theta}_{\mathbf{k}}^{(\mathcal{X})}$ and $\hat{\Theta}_{\mathbf{k}}^{(\mathcal{Y})}$ denote feature maps from source and target domains respectively. Following (Chen et al. 2018; Saito et al. 2019), we also adopt gradient reverse layer (GRL) (Ganin and Lempitsky 2015) to enable adversarial training where a GRL layer is placed between the domain classifier and the detection backbone network. During backpropagation, GRL will reverse the gradient that passes through from domain classifier to detection network.

For our instance-context alignment loss $\mathcal{L}_{\mathbf{ILoss}}$, we take the instance-level representation and context vector as inputs. The instance-level vectors are from RoI layer that each vector focuses on the representation of local object only. The context vector is from our proposed sub-network that combines hierarchical global features. We concatenate instance features with same context vector. Since context information is fairly different from objects, joint training detection and context networks will mix the critical information from each part, here we proposed a better solution that uses detach strategy to update the gradients. We will introduce it with details in the next section. Aligning instance and context representation simultaneously can help to alleviate the variances of object appearance, part deformation, object size, etc. in instance vector and illumination, scene, etc. in context vector. We define d_i as the domain label of i-th training image where $d_i = 1$ for the source and $d_i = 0$ for the target, so the instance-context alignment loss can be further formulated as:

$$\begin{aligned}\mathcal{L}_{\mathbf{ILoss}} = &-\frac{1}{N_{\mathcal{X}}}\sum_{i=1}^{N_{\mathcal{X}}}\sum_{i,j}(1 - d_i)\log \mathbf{P}_{(i,j)} \\ &-\frac{1}{N_{\mathcal{Y}}}\sum_{i=1}^{N_{\mathcal{Y}}}\sum_{i,j} d_i \log\left(1 - \mathbf{P}_{(i,j)}\right)\end{aligned} \tag{8}$$

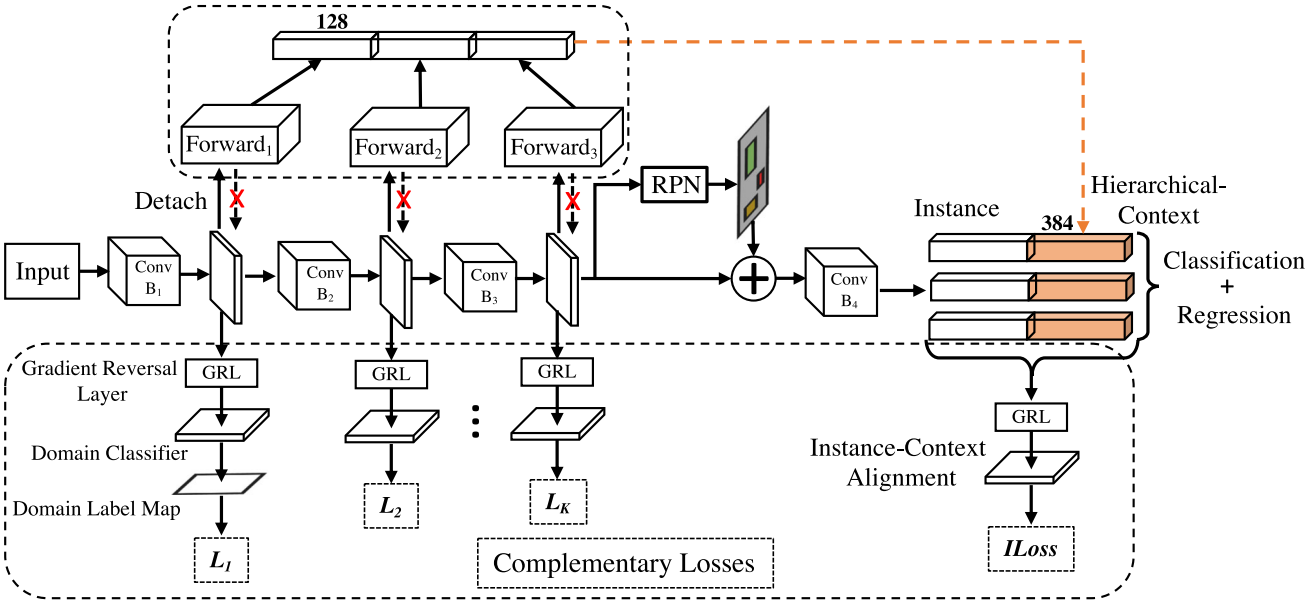

Fig. 9 Overview of our domain adaptive object detection framework. "RPN" is the region proposal network proposed in Faster RCNN (Ren et al. 2015) for generating object proposals. "GRL" is the gradient reverse layer (GRL) (Ganin and Lempitsky 2015) that the sign of the gradient will be reversed by passing through the GRL layer to optimize the base network, and the conventional gradient descent is applied for training the domain classifiers at different layers. More details please refer to Sect. 5

where $N_{\mathcal{X}}$ and $N_{\mathcal{Y}}$ denote the numbers of source and target examples. $\mathbf{P}_{(i,j)}$ is the output probabilities of the instance-context domain classifier for the j-th region proposal in the i-th image. So our total **SCL** (stacked complementary losses) objective $\mathcal{L}_{\mathbf{SCL}}$ can be written as:

$$\mathcal{L}_{\mathbf{SCL}} = \sum_{\mathbf{k}=1}^{\mathbf{K}} \mathcal{L}_{\mathbf{k}} + \mathcal{L}_{\mathbf{ILoss}} \tag{9}$$

Algorithm 1: Backward Pass of Our Detach Algorithm

INPUT: $\mathbf{G_c}$ is gradient of context network, $\mathbf{G_d}$ is the gradient of detection network, $\mathcal{L}_{det}$ is the detection objective, $\mathcal{L}_{\mathbf{SCL}}$ is the complementary objective;
for $t \leftarrow 1$ *to* n_{train_steps} **do**
  1. Update context net by detection and instance-context objectives: $\mathcal{L}_{det}$(w/o $\mathcal{L}_{rpn}$)+$\mathcal{L}_{\mathbf{ILoss}}$
  2. $\mathbf{G_d} \leftarrow$ stop-gradient($\mathbf{G_c}$;$\mathcal{L}_{det}$)
  3. Update detection net by detection and complementary objectives: $\mathcal{L}_{det}$+$\mathcal{L}_{\mathbf{SCL}}$

5.2 Framework Overall

Our detection part is based on the Faster RCNN (Ren et al. 2015), including the Region Proposal Network (RPN) and other modules. This is a conventional practice in many adaptive detection works. The objective of the detection loss is summarized as:

$$\mathcal{L}_{det} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{10}$$

where $\mathcal{L}_{cls}$ is the classification loss and $\mathcal{L}_{reg}$ is the bounding-box regression loss. To train the whole model using SGD, the overall objective function in the model is:

$$\min_{\mathcal{F},\mathbf{R}} \max_{\mathbf{D}} \mathcal{L}_{det}(\mathcal{F}(\mathbf{Z}), \mathbf{R}) - \lambda \mathcal{L}_{\mathbf{SCL}}(\mathcal{F}(\mathbf{Z}), \mathbf{D}) \tag{11}$$

where λ is the trade-off coefficient between detection loss and our complementary loss. $\mathbf{R}$ denotes the RPN and other modules in Faster RCNN.

6 Experiments and Analysis

6.1 Instance-level Image-to-image Translation

We conduct experiments on our collected dataset (CDTD). We also use COCO dataset (Lin et al. 2014) to verify the effectiveness of data augmentation.

Implementation Details. Our implementation is based on MUNIT[3] with PyTorch (Paszke et al. 2017). For I2I translation, we resize the short side of images to 360 pixels due

[3] https://github.com/NVlabs/MUNIT.

Table 4 Diversity scores on our dataset. We use the average LPIPS distance (Zhang et al. 2018) to measure the diversity of generated images

Method	Diversity			
	Sunny → Night	Sunny→Rainy	Sunny→Cloudy	Average
UNIT (Liu et al. 2017)	0.067	0.062	0.068	0.066
CycleGAN (Zhu et al. 2017)	0.016	0.008	0.011	0.012
MUNIT (Huang et al. 2018)	0.292	0.239	0.211	0.247
DRIT (Lee et al. 2018)	0.231	0.173	0.166	0.190
INIT w/ D_s	0.330	0.267	0.224	0.274
INIT w/o D_s	0.324	0.238	0.177	0.246
Real Images	0.573	0.489	0.465	0.509

Table 5 Comparison of Conditional Inception Score (CIS) and Inception Score (IS). To obtain high CIS and IS scores, a model is required to synthesis images that are more realistic, diverse with high-quality

	CycleGAN (Zhu et al. 2017)		UNIT (Liu et al. 2017)		MUNIT (Huang et al. 2018)		DRIT (Lee et al. 2018)		INIT w/ D_s		INIT w/o D_s	
	CIS	IS	CIS	IS	CIS	IS	CIS	IS	CIS	IS	CIS	IS
sunny→night	0.014	1.026	0.082	1.030	1.159	1.278	1.058	1.224	1.060	1.118	1.083	1.120
night→sunny	0.012	1.023	0.027	1.024	1.036	1.051	1.024	1.099	1.045	1.080	1.024	1.104
sunny→rainy	0.011	1.073	0.097	1.075	1.012	1.146	1.007	1.207	1.036	1.152	1.034	1.146
rainy→sunny	0.010	1.090	0.014	1.023	1.055	1.102	1.028	1.103	1.060	1.119	1.059	1.124
sunny→cloudy	0.014	1.097	0.081	1.134	1.008	1.095	1.025	1.104	1.040	1.142	1.025	1.147
cloudy→sunny	0.090	1.033	0.219	1.046	1.026	1.321	1.046	1.249	1.016	1.460	1.006	1.363
Average	0.025	1.057	0.087	1.055	1.032	1.166	1.031	1.164	**1.043**	**1.179**	1.039	1.167

The bold numbers denote the best results when compared with baselines, results of different settings, or other state-of-the-art methods

Table 6 Mask-RCNN with ResNet-50-FPN (Lin et al. 2017) detection and segmentation results on MS COCO 2017 `val` set

COCO 2017 training		COCO 2017 validation		Object detection (%)			Instance segmentation (%)		
Real	Synthetic	Real	Synthetic	Avg. Precision, IoU			Avg. Precision, mask		
				0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
✓		✓		37.7	59.2	40.8	34.3	56.0	36.2
✓			✓	30.4	49.7	32.6	27.8	46.6	29.2
	✓	✓		30.0	50.0	31.6	27.2	46.5	28.0
	✓		✓	30.5	49.7	32.7	27.8	46.4	29.0
✓	✓		✓	$32.6^{\uparrow 2.1}$	$52.6^{\uparrow 2.9}$	$34.2^{\uparrow 1.5}$	$29.0^{\uparrow 1.2}$	$49.0^{\uparrow 2.6}$	$29.8^{\uparrow 0.8}$
✓	✓	✓		$38.8^{\uparrow 1.1}$	$60.2^{\uparrow 1.0}$	$42.5^{\uparrow 1.7}$	$35.2^{\uparrow 0.9}$	$57.0^{\uparrow 1.0}$	$37.4^{\uparrow 1.2}$

to the limitation of GPU memory. For COCO image synthesis, since the training images (INIT dataset) and target images (COCO) are in different distributions, we keep the original size of our training image and crop 360×360 pixels to train our model, in order to learn more details of images and objects, meanwhile, ignore the global information. In this circumstance, we build our object part as an independent branch and each object is resized to 120×120 pixels during training. We set the trade-off hyper-parameters to $\lambda_g = 10$, $\lambda_o = 10$, $\lambda_{c_g} = 1$, $\lambda_{c_p} = 1$, $\lambda_{s_g} = 1$, $\lambda_{s_o} = 1$ following MUNIT (Huang et al. 2018).

6.1.1 Baselines

We perform our evaluation on the following four recent proposed state-of-the-art unpaired I2I translation methods:

– CycleGAN (Zhu et al. 2017): CycleGAN contains two translation functions ($\mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{X} \leftarrow \mathcal{Y}$), and the corresponding adversarial loss. It assumes that the input images can be translated to another domain and then can be mapped back with a cycle consistency loss.

– UNIT (Liu et al. 2017): The UNIT method is an extension of CycleGAN (Zhu et al. 2017) that is based on the shared latent space assumption. It contains two VAE-GANs and also uses cycle-consistency loss (Zhu et al. 2017) for learning models.
– MUNIT (Huang et al. 2018): MUNIT consists of an encoder and a decoder for each domain. It assumes that the image representation can be decomposed into a domain-invariant content space and a domain-specific style space. The latent vectors of each encoder are disentangled to a content vector and a style vector. I2I translation is performed by swapping content-style pairs.
– DRIT (Lee et al. 2018): The motivation of DRIT is similar to MUNIT. It consists of content encoders, attribute encoders, generators and domain discriminators for both domains. The content encoder maps images into a shared content space and the attribute encoder maps images into a domain-specific attribute space. A cross-cycle consistency loss is adopted for performing I2I translation.

6.1.2 Evaluation

We adopt the same evaluation protocol from previous unsupervised I2I translation works and evaluate our method with the LPIPS Metric (Zhang et al. 2018), Inception Score (IS) (Salimans et al. 2016) and Conditional Inception Score (CIS) (Huang et al. 2018).

LPIPS Metric. Zhang et al. proposed LPIPS distance (Zhang et al. 2018) to measure the translation diversity, which has been verified to correlate well with human perceptual psychophysical similarity. Following (Huang et al. 2018), we calculate the average LPIPS distance between 19 pairs of randomly sampled translation outputs from 100 input images of our test set. Following (Huang et al. 2018) and recommended by Zhang et al. (2018), we also use the pre-trained AlexNet (Krizhevsky et al. 2012) to extract deep features.

Results are summarized in Table 4, "INIT w/ D_s" denotes we train our model with shared discriminator between entire image and object. "INIT w/o D_s" denotes we build separate discriminators for image and object. Thanks to the coarse and fine styles we used, our average INIT w/ D_s score outperforms MUNIT with a notable margin. We also observe that our dataset (real image) has a very large diversity score, which indicates that the dataset is diverse and challenging.

Inception Score (IS) and Conditional Inception Score (CIS). We use the Inception Score (IS) (Salimans et al. 2016) and Conditional Inception Score (CIS) (Huang et al. 2018) to evaluate our learned models. IS measures the diversity of all output images and CIS measures diversity of output conditioned on a single input image, which is a modified IS that is more suitable for evaluating multimodal I2I translation task. The detailed definition of CIS can be referred to Huang et al. (2018). We also employ with Inception V3 model (Szegedy et al. 2016) to fine-tune our classification model on four domain category labels of our dataset. Other settings are the same as Huang et al. (2018). It can be seen in Table 5 that our results are consistently better than the baselines MUNIT and DRIT.

Image Synthesis on Multiple Datasets The visualization of our synthetic images is shown in Fig. 17. The left group images are on COCO and the right are on Cityscapes. We

Fig. 10 Visualization of our synthetic images. The left group images are from COCO and the right are from Cityscapes

Fig. 11 Visualization of multimodal results. We use randomly sampled style codes to generate these images and the darkness are slightly different across them

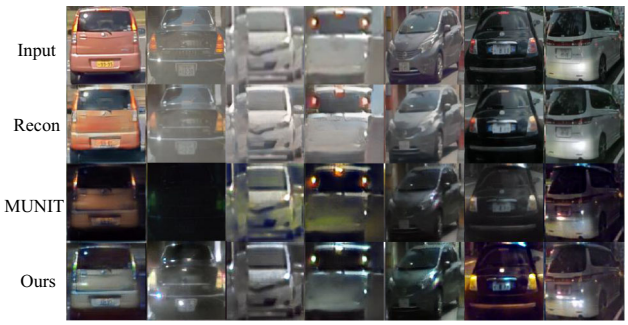

Fig. 12 Qualitative comparison on randomly selected instance level results. The first row shows the input objects. The second row shows the self-reconstruction results. The third and fourth rows show outputs from MUNIT and ours, respectively

Table 7 Improvement comparison on COCO detection with different image synthetic methods

COCO 2017 (%)	IoU	$IoU_{0.5}$	$IoU_{0.75}$
+Syn. (MUNIT Huang et al. 2018)	+0.7	+0.4	+1.0
+Syn. (Ours)	+1.1	+1.0	+ 1.7

Table 8 Performance decline when training and testing on real image, and comparing to results on synthetic image

	Metric	Percentage (%)
COCO	Det.&Seg.	↓19.1 & ↓19.0
Cityscapes	mIoU&mAcc	↓ **2.6** & ↓2.4

The bold number denotes the best results when compared with baselines, results of different settings, or other state-of-the-art methods
We adopt PSPNet (Zhao et al. 2017) with ResNet-50 (He et al. 2016) on Cityscapes (Cordts et al. 2016) and obtain (real&real): mIoU: 76.6%, mAcc: 83.1%; (syn.&syn.): 74.6%/81.1%

observe that the most challenging problem for multiple datasets synthesis is the inter-class variance among them.

6.2 Data Augmentation for Detection & Segmentation on COCO

We use Mask RCNN (He et al. 2017) framework for the experiments. A synthetic copy of entire COCO dataset is generated by our sunny→night model. We employ open-source implementation of Mask RCNN[4] for training the COCO models. For training, we use the same number of training epochs and other default settings including the learning rate schedule, #batchsize, etc.

All results are summarized in Table 6, the first column (group) shows the training data we used, the second group shows the validation data where we tested on. The third and fourth groups are detection and segmentation results, respectively. We can observe that our real-image trained model can obtain 30.4% mAP on synthetic validation images, this indicates that the distribution differences between original COCO and our synthetic images are not very huge. It seems that our generation process is more likely to do photo-metric distortions or brightness adjustment of images, which can be regarded as a data augmentation technique and has been verified the effectiveness for object detection in Liu et al. (2016). From the last two rows we can see that not only the synthetic images can help improve the real image testing performance, but the real image can also boost the results of synthetic images (both train and test on synthetic images). We also compare improvement with different generation methods in Table 7. The results show that our object branch can bring more benefits for detection task than the baseline. We also believe that the proposed data augmentation method can benefit to some limited training data scenarios like learning detectors from scratch (Shen et al. 2017; Law and Deng 2018; He et al. 2019; Duan et al. 2019).

We further conduct scene parsing on Cityscapes (Cordts et al. 2016). However, we didn't see obvious improvement in this experiment. Using PSPNet (Zhao et al. 2017) with ResNet-50 (He et al. 2016), we obtain mIoU: 76.6%, mAcc: 83.1% when training and testing on real images and 74.6%/81.1% on both synthetic images. We can see that the gaps between real and synthetic image are really small. We conjecture this case (no gain) is because the synthetic Cityscapes is too close to the original one. We compare the performance decline in Table 8. Since the metrics are different in COCO and Cityscapes, we use the relative percentage for comparison. The results indicate that the synthetic images may be more diverse for COCO since the decline is much smaller on Cityscapes.

6.2.1 Analysis

Qualitative Comparison We qualitatively compare our method with baseline MUNIT (Huang et al. 2018). Fig. 13 shows example results on sunny→night.

We randomly select one output for each method. It's obvious that our results are much more realistic, diverse with higher quality. If the object area is small, MUNIT (Huang et al. 2018) may fall into mode collapse and brings small artifacts around object area, in contrast, our method can overcome this problem through instance-level reconstruction. We also visualize the multimodal results in Fig. 11 with randomly sampled style vectors. It can be observed that the various degrees of darkness are generated across these images.

Instance Generation The results of generated instances are shown in Fig. 12, our method can generate more diverse objects (columns 1, 2, 6), more details (columns 5, 6, 7) with even the reflection (column 7). MUNIT sometimes fails to generate desired results if the global style is not suitable for the target object (column 2).

[4] https://github.com/facebookresearch/maskrcnn-benchmark.

Fig. 13 Case-by-case comparison on sunny→**night.** The first row shows the input images. The second and third rows show random outputs from MUNIT (Huang et al. 2018) and ours, respectively

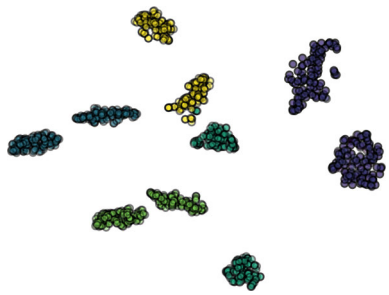

Fig. 14 Visualization of style distribution by t-SNE (Maaten and Hinton 2008). The groups with the same color are paired object and global styles of same domain

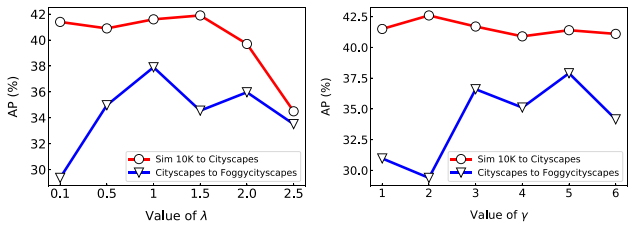

Fig. 15 Parameter sensitivity for the value of λ (left) and γ (right) in adaptation from Cityscapes to FoggyCityscapes and from Sim10k (Johnson-Roberson et al. 2016) to Cityscapes

Comparison of Local (Object) and Global Style Code Distributions. To further verify our assumption that the object and global styles are distinguishable enough to disentangle, we visualize the embedded style vectors from our w/ D_s model. The visualization is plotted by t-SNE tool (Maaten and Hinton 2008). We randomly sample 100 images and objects in the test set of each domain, results are shown in Fig. 14. The same color groups represent the paired global images and objects in the same domain. We can observe that the style vectors of same domain global and object images are grouped and separate with a remarkable margin, meanwhile, they are neighboring in the embedded space. This is reasonable and demonstrates the effectiveness of our learning process.

6.3 Domain Adaptive Object Object

Implementation Details. In all experiments, we resize the shorter side of the image to 600 following (Ren et al. 2015; Saito et al. 2019) with ROI-align (He et al. 2017). We train the model with SGD optimizer and the initial learning rate is set to 10^{-3}, then divided by 10 after every 50,000 iterations. Unless otherwise stated, we set λ as 1.0 and γ as 5.0, and we use $\mathbf{K} = 3$ in our experiments (the analysis of hyper-parameter $\mathbf{K}$ is shown in Table 11). We report mean average precision (mAP) with an IoU threshold of 0.5 for evaluation. Following (Chen et al. 2018; Saito et al. 2019), we feed one labeled source image and one unlabeled target one in each mini-batch during training. Our method is implemented on PyTorch platform.

6.3.1 Baselines and Our Results

The baselines and our results are shown in Tables 9 and 10. Following translation settings, we conduct experiments on three domain pairs: sunny→night (s2n), sunny→rainy (s2r) and sunny→cloudy (s2c). Since the training images in rainy domain are much fewer than sunny, for s2r experiment we randomly sample the training data in sunny set with the same number of rainy set and then train the detector. It can be observed that our method is consistently better than the baseline methods. We did not provide the results of s2c (faster) as we found that cloudy images are too similar to sunny in this dataset (nearly the same), thus the non-adapted result is very close to the adapted methods. Our code for domain adaptive object detection is available at: https://github.com/harsh-99/SCL.

Table 9 Adaptive detection results on our CDTD dataset

		Car	Sign	Person	mAP
s2n	Faster (Chen et al. 2018)	63.33	63.96	32.00	53.10
	Strong-Weak (Saito et al. 2019)	67.43	64.33	**32.53**	54.76
	Ours	**67.92**	**65.89**	32.52	**55.44**
	Ours+INIT	**69.72**	**66.87**	**33.87**	**56.80**
	Oracle	80.12	84.68	44.57	69.79
s2r	Faster (Chen et al. 2018)	70.20	72.71	36.22	59.71
	Strong-Weak (Saito et al. 2019)	**71.56**	78.07	39.27	62.97
	Ours	71.41	**78.93**	**39.79**	**63.37**
	Ours+INIT	**71.70**	**79.23**	**41.52**	**64.15**
	Oracle	71.83	79.42	45.21	65.49
s2c	Faster (Chen et al. 2018)	–	–	–	–
	Strong-Weak (Saito et al. 2019)	**71.32**	72.71	43.18	62.40
	Ours	71.28	**72.91**	**43.79**	**62.66**
	Ours+INIT	**73.13**	**72.98**	**44.82**	**63.64**
	Oracle	76.60	76.72	47.28	66.87

The bold numbers denote the best results when compared with baselines, results of different settings, or other state-of-the-art methods
"Ours+INIT" indicates that we train the INIT translation model first then generate the target domain images, after that, we train the gradient detach detection model with the generated and original data

Table 10 More adaptive detection results on other translation of the CDTD dataset

		Car	Sign	Person	mAP
n2c	Baseline (Chen et al. 2018)	61.36	53.91	34.35	49.87
	Ours	**62.21**	**55.21**	**36.6**	**51.34**
n2r	Baseline (Chen et al. 2018)	60.50	54.13	22.38	45.67
	Ours	**62.28**	**56.07**	**29.02**	**49.12**
n2s	Baseline (Chen et al. 2018)	60.06	50.43	33.47	48.12
	Ours	**60.73**	**56.09**	**35.2**	**50.67**
c2n	Baseline (Chen et al. 2018)	66.97	**62.44**	**38.43**	54.70
	Ours	**67.84**	62.06	35.73	**55.21**
c2r	Baseline (Chen et al. 2018)	71.04	70.84	38.01	59.96
	Ours	**71.10**	**77.47**	**38.01**	**62.19**
c2s	Baseline (Chen et al. 2018)	70.36	66.78	43.05	60.07
	Ours	**70.76**	**73.30**	**46.92**	**63.66**
r2c	Baseline (Chen et al. 2018)	66.75	57.11	32.71	52.19
	Ours	**68.93**	**63.27**	**34.02**	**55.41**
r2n	Baseline (Chen et al. 2018)	59.06	**50.36**	18.89	42.77
	Ours	**59.74**	49.99	**22.83**	**44.19**
r2s	Baseline (Chen et al. 2018)	**64.20**	55.22	32.13	50.52
	Ours	62.35	**57.20**	**36.89**	**52.15**

The bold numbers denote the best results when compared with baselines, results of different settings, or other state-of-the-art methods

Table 11 Analysis of hype-parameter **K** in stacked complementary losses

Method	**K**=2	**K**=3	**K**=4
from Cityscapes (Cordts et al. 2016) to Foggycityscapes (Sakaridis et al. 2018)	32.7	37.9	34.5
from PASCAL VOC (Everingham et al. 2010) to Clipart (Inoue et al. 2018)	39.0	41.5	39.3
from PASCAL VOC (Everingham et al. 2010) to Watercolor (Inoue et al. 2018)	54.7	55.2	53.4

Table 12 Ablation study (%) on Cityscapes to FoggyCityscapes (we use 150 m visibility, the densest one) adaptation

Method	Context	L_1	L_2	L_3	ILoss	Detach	AP on a target domain								
							Person	Rider	Car	Truck	Bus	Train	Mcycle	Bicycle	mAP
Faster (Non-adapted)							24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3
DA (CVPR'18)	✓						25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MAF (He and Zhang 2019) (ICCV'19)							28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SW (Saito et al. 2019) (CVPR'19)	✓						29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
Diversify (Kim et al. 2019) (CVPR'19)							30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
SW (re-impl. w/ VGG16)	✓						30.0	40.0	43.4	23.2	40.1	34.6	27.8	33.4	34.1
SW (re-impl. w/ Res101)	✓						29.1	41.2	43.8	26.0	43.2	27.0	26.2	30.6	33.4
Ours w/o Context	✗	*LS*	*FL*	✗	✗	✗	29.6	42.2	43.4	23.1	36.4	31.5	25.1	30.5	32.7
	✗	*LS*	*CE*	*FL*	✗	✗	29.0	41.4	43.9	24.6	46.5	28.5	27.0	32.8	34.2
	✗	*LS*	*CE*	*FL*	*FL*	✗	28.6	44.0	44.2	25.2	42.9	31.1	27.4	33.0	34.5
Ours w/ Context	✓	*LS*	*FL*	✗	✗	✗	28.5	42.6	43.8	23.2	41.6	24.9	28.3	30.3	32.9
	✓	*LS*	*FL*	*FL*	✗	✗	28.6	41.8	43.8	27.9	43.3	24.0	28.7	31.3	33.7
	✓	*LS*	*LS*	*FL*	✗	✗	28.8	**45.5**	44.3	28.6	44.6	29.1	27.8	31.4	35.0
	✓	*LS*	*CE*	*FL*	✗	✗	29.6	42.6	42.6	28.4	46.3	31.0	28.4	33.0	35.3
	✓	*LS*	*CE*	*FL*	✗	✓	30.0	42.7	44.2	30.0	**50.2**	34.1	27.1	32.2	36.3
	✓	*LS*	*FL*	*FL*	*FL*	✗	26.3	42.8	44.2	26.7	41.6	36.4	29.2	30.9	34.8
	✓	*LS*	*LS*	*FL*	*FL*	✓	29.5	43.2	44.2	27.0	42.1	33.3	29.4	30.6	34.9
	✓	*LS*	*FL*	*FL*	*FL*	✓	29.7	43.6	43.7	26.6	43.8	33.1	30.7	31.5	35.3
	✓	*LS*	*CE*	*FL*	*CE*	✗	28.3	41.9	43.1	25.4	45.1	35.5	26.7	31.6	34.7
	✓	*LS*	*CE*	*FL*	*FL*	✗	29.8	43.9	44.0	29.4	46.3	30.0	31.8	31.8	35.8
	✓	*LS*	*CE*	*FL*	*CE*	✓	29.0	42.5	43.9	28.9	45.7	42.4	26.4	30.5	36.2
	✓	*LS*	*CE*	*FL*	*FL*	✓	30.7	44.1	44.3	30.0	47.9	**42.9**	29.6	33.7	**37.9**
Ours w/ VGG16	✓	*LS*	*CE*	*FL*	*FL*	✓	**31.6**	44.0	**44.8**	**30.4**	41.8	40.7	**33.6**	**36.2**	**37.9**
Upper Bound (Saito et al. 2019)	–	–	–	–	–	–	33.2	45.9	49.7	35.6	50.0	37.4	34.7	36.2	40.3

The bold numbers denote the best results when compared with baselines, results of different settings, or other state-of-the-art methods
LS, Least-squares Loss; CE, Cross-entropy Loss; FL, Focal Loss; ILoss, Instance-Context Alignment Loss
The backbone network is ResNet-101

Fig. 16 Visualizations of our synthetic images on different source-target domain pairs. In each group, the first row is the reconstructed source domain images, and the second row is the synthetic target domain images

Fig. 17 More examples of our synthetic images on sunny→cloudy and sunny→rainy. Note that as the rainy images in our dataset look more like overcast weather with wet road, our results capture the attributes of training data very well

6.3.2 Ablation Results of Gradient Detach

To thoroughly verify the effectiveness of each component and generalization ability to other benchmarks of our proposed gradient detach method, we further investigate each component and design of our framework from Cityscapes (Cordts et al. 2016) to FoggyCityscapes (Sakaridis et al. 2018). Both source and target datasets have 2975 images in the training set and 500 images in the validation set. We design several controlled experiments for this ablation study. A consistent setting is imposed on all the experiments, unless when some components or structures are examined. In this study, we train models with the ImageNet (Deng et al. 2009) pre-trained ResNet-101 as a mainly used backbone, we also provide the results with pre-trained VGG16 model. We use four types of loss functions in SCL: LS: Least-squares Loss; CE: Cross-entropy Loss; FL: Focal Loss; ILoss: Instance-Context Alignment Loss.

Focal Loss (FL). Focal loss $\mathcal{L}_{\mathbf{FL}}$ (Lin et al. 2017) is adopted to ignore easy-to-classify examples and focus on those hard-to-classify ones during training:

$$\mathcal{L}_{\mathbf{FL}}\left(p_{\mathrm{t}}\right)=-f\left(p_{\mathrm{t}}\right)\log\left(p_{\mathrm{t}}\right),\ f\left(p_{\mathrm{t}}\right)=\left(1-p_{\mathrm{t}}\right)^{\gamma} \tag{12}$$

where $p_{\mathrm{t}}=p$ if $d_i=1$, otherwise, $p_{\mathrm{t}}=1-p$.

The results are summarized in Table 12. We present several combinations of four complementary objectives with their loss names and performance. We observe that "$LS|CE|FL|FL$" obtains the best accuracy with *Context* and *Detach*. It indicates that LS can only be placed on the low-level features (rich spatial information and poor semantic information) and FL should be in the high-level locations (weak spatial information and strong semantic information). For the middle location, CE will be a good choice. If you use LS for the middle/high-level features or use FL on the low-level features, it will confuse the network to learn hierarchical semantic outputs, so that *ILoss+detach* will lose effectiveness under that circumstance. This verifies that domain adaptive object detection relies heavily on the deep supervision, however, the diverse supervisions should be adopted in a controlled and correct manner. Furthermore, our proposed method performs much better than baseline Strong-Weak (Saito et al. 2019) (37.9% *vs.* 34.3%) and other state-of-the-arts.

Parameter Sensitivity on λ and γ. Figure 15 shows the results for parameter sensitivity of λ and γ in Eqs. 11 and 12. λ is the trade-off parameter between SCL and detection objectives and γ controls the strength of hard samples in *Focal Loss*. We conduct experiments on two adaptations: Cityscapes → FoggyCityscapes (blue) and Sim10K (Johnson-Roberson et al. 2016) → Cityscapes (red). On Cityscapes → FoggyCityscapes, we achieve the best performance when $\lambda=1.0$ and $\gamma=5.0$ and the best accuracy is 37.9%. On Sim10K → Cityscapes, the best result is obtained when $\lambda=0.1$, $\gamma=2.0$.

Hyper-parameter K Analysis. Table 11 shows the results for sensitivity of hyper-parameter **K** in Figure 9. This parameter controls the number of SCL losses and context branches. It can be observed that the proposed method performs best when $\mathbf{K}=3$ on all three datasets.

7 Conclusion

In this work, we have introduced a large-scale cross-domain dataset for the instance-level image-to-image translation and domain adaptive object detection tasks. We presented INIT method for instance-aware translation with unpaired training data. Extensive qualitative and quantitative results demonstrate that the proposed method can capture the details of objects and produce realistic and diverse images. We also addressed unsupervised domain adaptive object detection through a novel training strategy, gradient detach, for the convolutional neural networks. Our future work will focus on exploring the domain-shift tasks from scratch, i.e., without the pre-trained models (Shen et al. 2017, 2019; He et al. 2019; Zhu et al. 2019) to avoid involving bias from the pre-trained dataset.

References

Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., & Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*.

Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Finding tiny faces in the wild with generative adversarial network. In *CVPR*.

Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., & Yao, T. (2019). Exploring object relation in mean teacher for cross-domain detection. In *CVPR*.

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*

Cheung, B., Livezey, J. A., Bansal, A. K., & Olshausen, B.A. (2015). Discovering hidden factors of variation in deep networks. In *ICLR workshop*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Object detection with keypoint triplets. arXiv preprint arXiv:1904.08189.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*,*88*(2), 303–338.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *ICML*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural

networks. *The Journal of Machine Learning Research*, *17*(1), 2096–2030.
Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
Girshick, R. (2015). Fast R-CNN. In *ICCV*.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.
He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In: *Proceedings of the IEEE international conference on computer vision* (pp. 4918–4927).
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *ICCV*.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
He, Z., & Zhang, L. (2019). Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*.
Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.
Huang, X., & Belongie, S.J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.
Huang, X., Liu, M.Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *ECCV*.
Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, *36*(4), 1–14.
Inoue, N., Furuta, R., Yamasaki, T., & Aizawa, K. (2018). Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*.
Isola, P., Zhu, J.Y., Zhou, T., & Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition*.
Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., & Vasudevan, R. (2016). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983.
Karacan, L., Akata, Z., Erdem, A., & Erdem, E. (2016). Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:1612.00215.
Kim, T., Jeong, M., Kim, S., Choi, S., & Kim, C. (2019). Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*.
Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*.
Laffont, P. Y., Ren, Z., Tao, X., Qian, C., & Hays, J. (2014). Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics Proceedings of SIGGRAPH*, *33*(4), 1–11.
Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *ECCV*.
Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., & Yang, M.H. (2018). Diverse image-to-image translation via disentangled representations. In *ECCV*.
Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., & Lin, L. (2018). Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *2018 ACM multimedia conference on multimedia conference* (pp. 645–653). ACM.
Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., & Belongie, S.J. (2017). Feature pyramid networks for object detection. In *CVPR*.
Lin, T.Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *ICCV*.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.
Liu, M.Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NIPS*.
Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., & Berg, A.C. (2016). SSD: Single shot multibox detector. In *ECCV*.
Long, M., Zhu, H., Wang, J., & Jordan, M.I. (2016). Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*.
Ma, S., Fu, J., Chen, C.W., & Mei, T. (2018). Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*.
Maaten, L.v.d., & Hinton, G., (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.
Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., & LeCun, Y. (2016). Disentangling factors of variation in deep representation using adversarial training. In *NIPS*
Mechrez, R., Talmi, I., & Zelnik-Manor, L. (2018). The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, 768–783.
Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
Mo, S., Cho, M., & Shin, J. (2019). Instance-aware image-to-image translation. In *International conference on learning representations*. https://openreview.net/forum?id=ryxwJhC9YX.
Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., & Kim, K. (2018) Image to image translation for domain adaptation. In *CVPR*.
Nguyen, V., Vicente, Y., Tomas, F., Zhao, M., Hoai, M., & Samaras, D. (2017). Shadow detection with conditional generative adversarial networks. In *ICCV*.
Panareda Busto, P., & Gall, J. (2017). Open set domain adaptation. In *ICCV*.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS workshop*.
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536–2544).
Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1406–1415).
Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., & Saenko, K. (2017). Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924.
Peng, X., Usman, B., Saito, K., Kaushik, N., Hoffman, J., & Saenko, K. (2018). Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. arXiv preprint arXiv:1806.09755.
Radim Tyleček, R. Š. (2013). *Spatial pattern templates for recognition of objects with regular structure*. Saarbrucken, Germany: In Proceeding GCPR.
Raj, A., Namboodiri, V. P., & Tuytelaars, T. (2015). Subspace alignment based domain adaptation for rcnn detector. arXiv preprint arXiv:1507.05578.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*.
Saito, K., Ushiku, Y., Harada, T., & Saenko, K. (2019). Strong-weak distribution alignment for adaptive object detection. In *CVPR*.
Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, *126*(9), 973–992.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *NIPS*.
Sangkloy, P., Lu, J., Fang, C., Yu, F., & Hays, J. (2017). Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*.
Shen, Z., He, Z., & Xue, X. (2019). Meal: Multi-model ensemble via adversarial learning. In *AAAI*.
Shen, Z., Huang, M., Shi, J., Xue, X., & Huang, T. (2019). Towards instance-level image-to-image translation. In *CVPR*.
Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., & Xue, X. (2017). Dsod: Learning deeply supervised object detectors from scratch. In *ICCV*.
Shen, Z., Liu, Z., Li, J., Jiang, Y. G., Chen, Y., & Xue, X. (2019). Object detection from scratch with deep supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 398–412.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.
Tzeng, E., Burns, K., Saenko, K., & Darrell, T. (2018). Splat: Semantic pixel-level adaptation transforms for detection. arXiv preprint arXiv:1812.00929.
Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *CVPR*.
Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *(IEEE) conference on computer vision and pattern recognition (CVPR)*.
Wang, X., Cai, Z., Gao, D., & Vasconcelos, N. (2019). Towards universal object detection by domain attention. In *CVPR*.
Wu, Y., Winston, E., Kaushik, D., & Lipton, Z. (2019). Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*.
Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.
Zhang, R., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.
Zhang, Z., Yang, L., & Zheng, Y. (2018). Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *CVPR*.
Zhao, H., Des Combes, R. T., Zhang, K., & Gordon, G. (2019). On learning invariant representations for domain adaptation. In *ICML*.
Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *CVPR*.
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*.
Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017). Toward multimodal image-to-image translation. In *Advances in neural information Processing Systems*.
Zhu, R., Zhang, S., Wang, X., Wen, L., Shi, H., Bo, L., & Mei, T. (2019). Scratchdet: Training single-shot object detectors from scratch. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2268–2277).