CROSS-DOMAIN AUTONOMOUS DRIVING PERCEPTION USING CONTRASTIVE APPEARANCE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Addressing domain shifts for complex perception tasks in autonomous driving has long been a challenging problem. In this paper, we show that existing domain adaptation methods pay little attention to the *content mismatch* issue between source and target images, thereby weakening the domain adaptation performance and the decoupling of domain-invariant and domain-specific representations. To solve the aforementioned problems, we propose an image-level domain adaptation framework that aims at adapting source-domain images to the target domain with content-aligned image pairs. Our framework consists of three mutual-beneficial modules in a cycle: a *cross-domain content alignment* module to generate source-target pairs with consistent content representations in a self-supervised manner, *a reference-guided image synthesis* using the generated content-aligned source-target image pairs, and a *contrastive learning* module to self-supervise domain-invariant feature extractor from the generated images. Our contrastive appearance adaptation is task-agnostic and robust to complex perception tasks in autonomous driving. Our proposed method demonstrates state-of-the-art results in cross-domain object detection, semantic segmentation, and depth estimation as well as better image synthesis ability qualitatively and quantitatively.

1 INTRODUCTION

Building scalable and robust perception capabilities such as object detection (Chen et al., 2018; Faster, 2015; He & Zhang, 2019) and semantic segmentation (Chen et al., 2017; Xie et al., 2021) is a challenging task in autonomous driving systems. A fundamental challenge is the domain shift, where the systems are expected to work in various conditions such as under adverse weather, different illuminations, and varying geographic locations. In theory, supervised learning can be used to train object detectors and semantic segmentation models with data and labels acquired in various conditions, but in practice, this approach is too expensive due to the high cost of data acquisition and annotation, the countless variants of the road conditions, and some potential changes in hardware, sensors, and simulation environments in autonomous driving.

To mitigate such issues, a practical solution is domain adaptation, which aims at adapting a model trained with labels in a source domain to a novel target domain without labels. Recent domain adaptation methods mainly focus on the feature-level adaptation, including discrepancy-based (Long et al., 2015; 2016), adversarial feature learning (Chen et al., 2018), self-training (RoyChowdhury et al., 2019; Khodabandeh et al., 2019) and knowledge distillation (He et al., 2022; Deng et al., 2021; Li et al., 2022b; Cai et al., 2019). Beyond these methods, image-level adaptation (Zhu et al., 2017; Choi et al., 2018) can reduce appearance differences between the two domains by generating target-like images from the source images. An inherent advantage of image-level adaptation is that it is task-agnostic, which means that the generated target-like images can be used for a wide variety of downstream tasks, making it highly suitable for multi-task scenarios such as autonomous driving. However, existing image-level adaptation methods tend to suffer from visual artifacts caused by imperfect image synthesis, degrading overall domain adaptation performance.

In this paper, we hypothesize that the inferior performance of image-level adaptation is attributed to the *content mismatches* between samples from the source and target domains. We define content mismatches by the limited semantic correspondences or layout similarity (Li et al., 2020b) between a source-domain image and a target-domain image. We provide a motivating example in Figure 1. In this example, we aim to translate an image x_s from the source domain $\mathbb{S}$ to the target domain $\mathbb{T}$ by using a reference image x_t^1 or x_t^2 from $\mathbb{T}$. We observe that reference x_t^1 shares more similar semantic

Figure 1: Addressing content mismatches between the source image and the reference image is key to effective image synthesis and domain adaptation. This example shows that with the same source image, choosing a reference image with well-aligned semantic correspondences leads to better quality in image synthesis. Best viewed in color.

correspondences (green region) with the source image than reference x_t^2 (red region). Therefore, the generated image $\tilde{x}_t^1$ has better image quality and fewer visual artifacts than $\tilde{x}_t^2$. This example motivates that choosing the right reference to reduce content mismatches is key to achieving effective image translation and domain adaptation. Unfortunately, few existing domain adaptation methods consider such content mismatches between samples from the source and target domains. It has been shown that addressing such semantic correspondences enables the disentanglement of the domain-invariant and domain-specific representations (Zhou et al., 2021).

To better model these semantic correspondences, in this paper, we propose a novel framework for image-level domain adaptation that incorporates cross-domain content alignment (CDCA), contrastive learning (Chen et al., 2020), and reference-guided image synthesis. Our method is general-purpose and task-agnostic, which can support different perception tasks such as object detection, semantic segmentation, and depth estimation. The modules in our framework are mutual-beneficial. The cross-domain content alignment (CDCA) module builds pairs of images from the source and the target domain, respectively, such that the discrepancy in the content representation in each image pair is minimized. Such content representation can be extracted from a domain-invariant feature extractor trained by contrastive learning on source images and generated images from the reference-guided image synthesis module. Specifically, our reference-guided image synthesis takes a source image and a reference image from the target domain as input and synthesizes a new image that fuses the content of the source with the style of the reference. We term this new image a *target-like* image. The source image and the generated target-like image form a pair of augmented views that can be used in contrastive learning to learn a feature extractor to output domain-invariant features. The feature extractor can then be used by the CDCA module to retrieve a better content-aligned reference image that subsequently boosts the overall performance. The three modules mutually improve each other and eventually converge to an accurate adaptation.

We have conducted experiments of our proposed method with multiple downstream tasks for cross-domain perception in autonomous driving including object detection, semantic segmentation, and depth estimation. Our results show that our method can deal with domain shifts effectively and outperform all existing state-of-the-art methods by a large margin. Our contributions are:

- A novel task-agnostic image-level domain adaptation method that addresses the content mismatches between the source domain and the target domain by using reference-guided image synthesis and contrastive learning.
- Ablation studies and result analysis that explain the merits of our method in modeling the implicit semantic correspondences between domains, resulting in the better disentanglement of the domain-invariant and domain-specific knowledge.
- Extensive experiments that demonstrate the effectiveness of our method on multiple datasets in autonomous driving for multiple tasks including cross-domain object detection, semantic segmentation, and depth estimation, achieving state-of-the-art results.

2 Related Work

Cross-domain perception systems. Domain adaptation techniques for specific perception tasks such as object detection and semantic segmentation have been developed based on the main principles originally developed for unsupervised domain adaptation for visual data (Csurka, 2017; Zhao et al., 2020b; Oza et al., 2021) such as discrepancy-based methods (Long et al., 2015; 2016; Kang et al., 2019), adversarial training (Nam et al., 2021; Chen et al., 2022), self-training (Sun et al., 2019;

Prabhu et al., 2021) and knowledge distillation (Xu et al., 2021; Nguyen-Meidine et al., 2021). Adversarial training was pioneered in object detection by reducing the domain discrepancy in a min-max manner with a domain classifier (Chen et al., 2018; Hoffman et al., 2018). To improve the performance, the gradient reversal layer and domain classifier (Goodfellow et al., 2020) were designed for adversarial feature learning. Similarly, one can also minimize the domain discrepancy by learning the domain-invariant representations for semantic segmentation (Vu et al., 2019). However, solely using adversarial learning in the feature space only achieved marginal accuracy improvement in complex scenarios. Self-training algorithms (Zou et al., 2018; RoyChowdhury et al., 2019) utilize the supervision from the pre-trained model in the source domain for retraining in the target domain. However, these methods suffer from the low quality of generated pseudo labels in the target domain.

Recent knowledge distillation algorithms (Cai et al., 2019; Tian et al., 2021; Zhang et al., 2021; Deng et al., 2021) introduced the Mean-Teacher framework to the domain-adaptive object detection and semantic segmentation. MGADA (Zhou et al., 2022) conducted the multi-granularity (category-level, instance-level and pixel-level) domain adaptation for object detection. AT (Li et al., 2022b) aimed to improve the quality of the generated pseudo-label in the target domain through leveraging domain adversarial learning and mutual learning. DaFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b) achieved the current state-of-the-art domain adaptive semantic segmentation performance by introducing SegFormer (Xie et al., 2021) for more effective knowledge distillation. Recently, cross-domain perception tasks have also been solved by using multi-source domain adaptation (Yao et al., 2021; Peng et al., 2019; Wu et al., 2022) that aims to preserve the target-relative knowledge from various source domains to promote the detection performance in the target domain (Wu et al., 2022). While being successful for some scenarios, these domain adaptation approaches remain limited when there exist significant content mismatches between the source and target domains.

Content-aware adaptation. Some efforts have been done to address content mismatches between the source and target images. CCM (Li et al., 2020b) constructed the positive pairs for better domain adaptation in the label space through pixel-wise similarity matching. SIGMA (Li et al., 2022a) proposed to design a bipartite graph matching to find well-matched instance pairs across graphs for reducing content mismatches. Recent works (Sakaridis et al., 2019; 2020; 2021) propose to construct normal-adverse image pairs that have a similar layout based on the GPS information. The utilization of the normal images collected with good visibility results in better domain adaptation performance (Bruggemann et al., 2023; Sakaridis et al., 2020), which also demonstrates that aligned content representation can boost the domain adaptation performance. Compared to these works, our work focuses on a more generic cross-domain content alignment without leveraging annotation labels. We support multiple downstream tasks by building our method upon image synthesis methods.

Image synthesis for task-agnostic domain adaptation. Image synthesis methods can learn to generate target-like images from images in the source and target domains to reduce the domain gap for task-agnostic image-level domain adaptation (Romera et al., 2019; Zheng et al., 2020b; Nam et al., 2021; Yu et al., 2022). State-of-the-art image translation methods are CycleGAN with the cycle-consistency loss (Zhu et al., 2017) and its variants (Choi et al., 2018; Zheng et al., 2020b; Pizzati et al., 2021). Recent works (Hoffman et al., 2018; Romera et al., 2019; Chen et al., 2020; Deng et al., 2021) introduced image translations for cross-domain image classification (Nam et al., 2021), semantic segmentation (Hoffman et al., 2018; Romera et al., 2019) and object detection (Deng et al., 2021; Yu et al., 2022). However, solely using the cycle-consistency loss cannot guarantee the disentanglement of the domain-invariant and domain-specific representations. Specifically, the generated images in the target domain may lose content representation or yield unnecessary visual artifacts. Recent reference-guided image synthesis (Huang et al., 2018; Lee et al., 2018; Zheng et al., 2020a; Jiang et al., 2020) can combine the *style* representation from a reference image in the target domain and the content representation from a source image to generate a target-like image, resulting in better feature disentanglement. However, these methods do not consider the content mismatches between the source image and the reference image, which might result in inferior disentanglement. In this work, we propose an effective approach that wires contrastive learning and image translation in a mutual-beneficial way for retrieving content-aware reference images, thereby reducing content mismatches and improving overall image translation and domain adaptation performance.

3 Our Method

Let us now formulate our problem by assuming the domain adaptation from a source domain $\mathbb{S}$ to a target domain $\mathbb{T}$, where their data distribution is different, i.e., $\mathcal{P}_{\mathbb{S}} \neq \mathcal{P}_{\mathbb{T}}$. The source domain

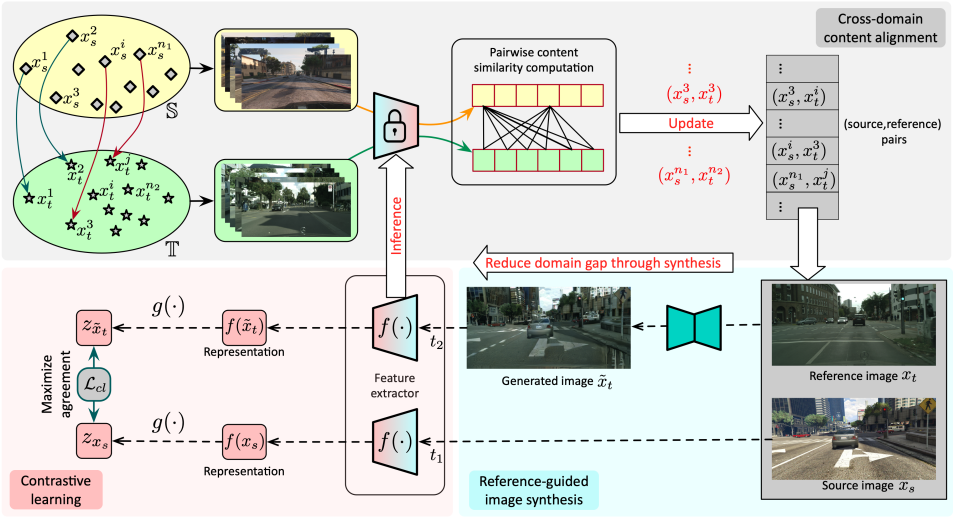

Figure 2: The core of our image-level domain adaptation is a cycle of three mutual-beneficial modules: cross-domain content alignment (CDCA), reference-guided image synthesis and contrastive learning. CDCA uses the domain-invariant feature extractor learned by contrastive learning to construct source-reference image pairs for training the reference-guided image synthesis module to produce target-like images. The source and target-like images can be regarded as augmented views for contrastive learning to improve the domain-invariant feature extractor. Final target-like images and source labels can be adopted for downstream perception tasks.

is labeled while the target domain is unlabeled. Let $\mathbb{S} = \{x_s^i, y_s^i\}$, where x_s^i is the source image and y_s^i is the corresponding annotation in the source domain for $i \in 1..N_s$. Similarly, $\mathbb{T} = \{x_t^i\}$ for $i \in 1..N_t$, where N_s and N_t indicate the number of images in each domain, respectively. For cross-domain perception tasks in autonomous driving, we assume that the labels for object detection are 2D bounding boxes and for semantic segmentation the labels are pixel-level annotations.

An overview of the proposed method for domain adaptation is shown in Figure 2, which mainly includes three modules in a cycle: 1) cross-domain content alignment (CDCA) to construct source-reference pairs between images from the source and target domain; 2) reference-guided image synthesis; and 3) contrastive learning for domain-invariant feature extraction. Based on the extracted content representation from both source and target images, the CDCA module retrieves the target images with the most consistent content representation with each source image and constructs the source-reference pairs. Given such source-reference pairs, reference-guided image-to-image translation generates images in the target domain for reducing the domain gap. Later, source images and the corresponding generated images are used in the contractive learning module for learning domain-invariant features, for further content similarity computation in CDCA.

3.1 Cross-domain Content Alignment

In each cycle, to perform cross-domain content alignment (CDCA), all the real source and target images from the whole training datasets are fed into the feature extractor $f(\cdot)$ to obtain the content representations. The architecture of $f(\cdot)$ is a ResNet-50 (He et al., 2016). $f(\cdot)$ is initialized with the pre-trained weights on ImageNet dataset and optimized by contrastive learning described in Section 3.3. To build source-reference image pairs, we first compute the content representation similarity between the source and the reference image in each pair by cosine similarity. For each source image, the target image with the highest similarity score is selected to form a source-reference image pair during the training procedure.

3.2 Reference-guided Image Synthesis

After performing CDCA, the constructed pairs are then used for the reference-guided image synthesis, which aims at generating target-like images based on the source images to reduce the domain gap. Let the source-reference image pair generated by CDCA be (x_s, x_t). Our reference-guided image synthesis is a dual-stream neural network that fuses the content of the source image and the style of the

reference image into a final image. Particularly, x_s is fed into the *content stream* to extract the source content feature, while x_t is put into the *style stream* to extract domain-specific representations (style or appearance). We use FAdaIN and FADE (Jiang et al., 2020) for borrowing the style representation from x_t and preserving the content representation of x_s. The two modules that can perform the feature adaptive normalization are constructed in a multi-scale manner to capture the coarse-to-fine content and style information. The generated image $\tilde{x}_t$ is target alike, which means that the pair $(x_s, \tilde{x}_t)$ has a smaller domain gap compared to the (x_s, x_t) image pair. To improve the photorealism of generated images we use a multi-scale discriminator (Huang et al., 2018). Our image translation is trained with hinge-based adversarial loss. The generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$ can be written as:

$$\mathcal{L}_G = -\mathbb{E}[D(\tilde{x}_t)] + \lambda_{fm}\mathcal{L}_{fm}(\tilde{x}_t, x_t), \tag{1}$$

$$\mathcal{L}_D = -\mathbb{E}[\min(-1 + D(x_t), 0)] - \mathbb{E}[\min(-1 - D(\tilde{x}_t), 0)], \tag{2}$$

where D is the discriminator; $\mathcal{L}_{fm}$ is the feature matching loss (Wang et al., 2018) to enforce the similarity of the intermediate feature representations at different layers of the multi-scale discriminators. Since the content representations of the source images can be better preserved with the aligned source-reference image pairs, we found that perceptual loss (Johnson et al., 2016) is no longer required, which can reduce the computation cost and alleviate memory usage.

3.3 Domain-invariant Representation Learning

The contrastive learning framework is adopted for projecting both the source and target images into the same feature space, optimizing the feature extractor $f(\cdot)$ for further use in CDCA. Particularly, we can view the source-generated image pair $(x_s, \tilde{x}_t)$ from the reference-guided image synthesis stage as augmented views of a latent domain-invariant representation, and therefore we can use contrastive learning to train the feature extractor to be insensitive to the domain gap between x_s and $\tilde{x}_t$. We feed the source image x_s and corresponding generated image $\tilde{x}_t$ into the feature extractor $f(\cdot)$ to obtain feature representations, and then pass these features to a projection head $g(\cdot)$ to obtain the final features z_{x_s} and $z_{\tilde{x}_t}$ for computing the contrastive loss. Similar to SimCLR (Chen et al., 2020), we apply transform operations from the transformation sets $\mathcal{T}$ (including *random resizing and cropping*, *color jitter* and *random greyscale*) to get augmented input data $x_s \leftarrow t_1(x_s)$ and $\tilde{x}_t \leftarrow t_2(\tilde{x}_t)$ where t_1, t_2 are augmentation operators drawn randomly from $\mathcal{T}$. The contrastive loss $\mathcal{L}_{cl}$ can be written as

$$\mathcal{L}_{cl} = -\log \frac{\exp(sim(z_{x_s}, z_{\tilde{x}_t}))/\tau)}{\sum_{x \in X, x \neq x_s} \exp(sim(z_{x_s}, z_x)/\tau)}, \tag{3}$$

where $sim(u, v) = u^T v/\|u\|\|v\|$ denotes the pair-wise content similarity between feature u and v and τ is the temperature parameter. X is the set of total images in the current mini-batch. Note that our contrastive learning does not utilize the original target images for training since there is no correspondence between source and target images for constructing positive pairs and we can already reduce the domain gap through image synthesis.

3.4 Cross-domain Perception

We train the modules in our framework in an end-to-end manner. Particularly, we integrate the reference-guided image synthesis and contrastive learning through joint training and iteratively optimize each module. In this work, we consider two supervised downstream tasks with object detection and semantic segmentation, and an unsupervised downstream task with depth estimation, respectively. For object detection and semantic segmentation, we adopt the target-like images and the source labels as training data for supervised learning. For depth estimation, we consider unsupervised monocular depth estimation and assume that a well-trained depth estimator is available for the source domain. We therefore perform domain adaptation from the target domain to the source domain, and apply the depth estimator on the generated source-like images.

4 Experiments

4.1 Implementation

The image resolution of the reference-guided image synthesis is set to 1024×512 for both cross-domain object detection and 2048×1024 for cross-domain semantic segmentation. We perform all the cross-domain object detection experiments based on the MMDetection (Chen et al., 2019). We adopt Faster R-CNN (Faster, 2015) with ResNet-50 (He et al., 2016) pre-trained on ImageNet as the backbone network. The shorter side of each input image is resized to 600 pixels. For

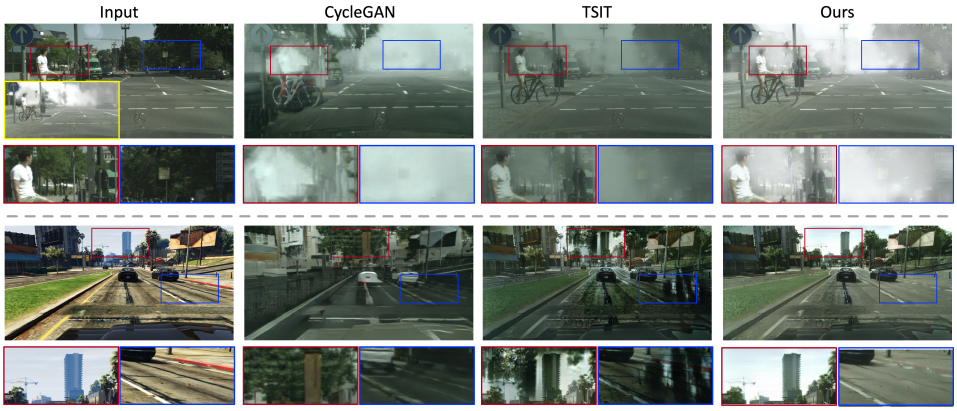

Figure 3: Generated image quality comparison between CycleGAN, TSIT and our method under Cityscapes→Foggy Cityscapes (above dashed line) and Sim10k→Cityscapes (below the dashed line). The ground truth from the Foggy Cityscapes dataset is provided in the yellow box for reference.

Table 1: FID scores of different image synthesis methods. Our method outperforms both previous non-reference method (CycleGAN) and reference-based method (TSIT).

Methods	CycleGAN (Zhu et al., 2017)	TSIT (Jiang et al., 2020)	Ours
Cityscapes→Foggy Cityscapes	25.76	7.35	**5.68**
Sim10k→Cityscapes	77.31	60.24	**48.23**

semantic segmentation, we adopt the DACS (Tranheden et al., 2021) DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b) as backbones and conduct the experiments following the official instructions. We choose both the source data and the target-like translated data to optimize the both detection and segmentation models, which encourages the training model without being biased. We provide more implementation details for each task in our appendix.

4.2 Evaluating Task-agnostic Image Synthesis

Let us first evaluate the quality of the generated images qualitatively and quantitatively as high-quality image synthesis could imply high performance for the downstream perception tasks. We compare our reference-guided image synthesis with two representative image synthesis methods namely CycleGAN (Zhu et al., 2017) and TSIT (Jiang et al., 2020). CycleGAN translates the source image to the target domain without reference. TSIT is a reference-guided image synthesis method similar to ours. For quantitatively measuring the image quality of the synthesized images, we adopt FID (Heusel et al., 2017) (lower is better) as the evaluation metric. We report the qualitative and quantitative results in Figure 3 and Table 1 respectively, under the adaptation Cityscapes→Foggy Cityscapes and Sim10k→Cityscapes. Our image synthesis outperforms both CycleGAN and TSIT on both datasets by a large margin. Such improvement can be explained by the improved source-reference image pairs obtained from the improved domain-invariant feature extractors trained by contrastive learning.

4.3 Evaluating Cross-Domain Object Detection

We first report results on Cityscapes → Foggy Cityscapes in Table 2. We report the average precision (AP) of 8 categories on the Foggy Cityscapes are computed as well as the mAP. We compare our method with the recent state-of-the-art methods including SCL (Shen et al., 2019), GPA (Xu et al., 2020b), UMT (Deng et al., 2021), MeGA-CDA (Vs et al., 2021), CDG (Li et al., 2021), MGADA (Zhou et al., 2022), SIGMA (Li et al., 2022a), TDD (He et al., 2022) and AT (Li et al., 2022b). As can be seen, our proposed method outperforms the existing cross-domain object detection methods by a large margin. The qualitative results are shown in Figure 4. As illustrated, the proposed method could accurately detect small objects under dense fog, e.g., the bicycle in the first row.

We then report **synthetic-to-real** and **cross-camera** detection results from Sim10k/KITTI to Cityscapes. Sim10k is a simulated dataset containing 10,000 images. KITTI is a scene dataset (7,481 labeled images) with a different camera setup as Cityscapes. The validation set of Cityscapes is used for evaluation. Only the category car is used for evaluation under both settings. The quantita-

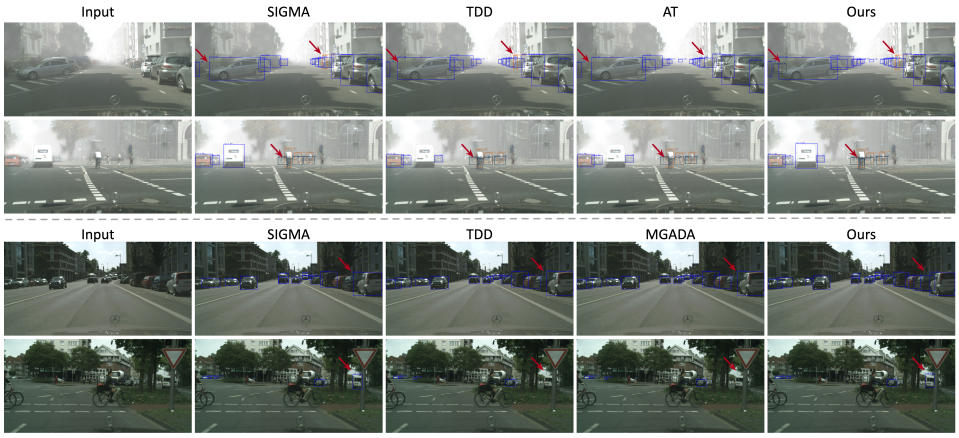

Figure 4: Qualitative comparisons of cross-domain object detection methods on Cityscapes→Foggy Cityscapes and Sim10k→Cityscapes. Models are trained on the car category.

Table 2: Cross-domain object detection on the Foggy Cityscapes dataset using Cityscapes→Foggy Cityscapes adaptation.

Methods	Detector	Backbone	person	rider	car	truck	bus	train	motor	bicycle	mAP↑
SCL (Shen et al., 2019)	F-RCNN	VGG-16	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
GPA (Xu et al., 2020b)	F-RCNN	ResNet-50	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
UMT (Deng et al., 2021)	F-RCNN	VGG-16	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
MeGA-CDA (Vs et al., 2021)	F-RCNN	VGG-16	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
CDG (Li et al., 2021)	F-RCNN	VGG-16	38.0	47.4	53.1	34.2	47.5	41.1	38.3	38.9	42.3
MGADA (Zhou et al., 2022)	F-RCNN	VGG-16	43.9	49.9	60.6	29.6	50.7	39.0	38.3	42.8	44.3
SIGMA (Li et al., 2022a)	F-RCNN	ResNet-50	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
TDD (He et al., 2022)	F-RCNN	ResNet-50	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	49.2
AT (Li et al., 2022b)	F-RCNN	VGG-16	45.5	55.1	64.2	35.0	56.3	**54.3**	38.5	**51.9**	50.9
Ours	F-RCNN	ResNet-50	**53.7**	**58.3**	**72.2**	**36.6**	**60.6**	51.3	**44.3**	51.6	**53.6**

tive results are reported in Table 3 and we provide the qualitative comparisons with SIGMA (Li et al., 2022a), TDD (He et al., 2022) and MGADA (Zhou et al., 2022) in Figure 4 for Sim10k→Cityscapes adaptation. We also perform multi-source cross-domain object detection in our appendix.

Table 3: Synthetic-to-real/Cross-camera domain adaptation from Sim10k/KITTI to Cityscapes on object detection task.

Methods	Detector	Backbone	mAP (*car*) Sim10k / KITTI ↑
CST (Zhao et al., 2020a)	F-RCNN	VGG-16	44.5 / 43.6
MeGA-CDA (Vs et al., 2021)	F-RCNN	VGG-16	44.8 / 43.0
UMT (Deng et al., 2021)	F-RCNN	VGG-16	43.1 / -
CDN (Su et al., 2020)	F-RCNN	VGG-16	49.3 / 44.9
CFA (Hsu et al., 2020)	FCOS	VGG-16	49.0 / 43.2
CFA (Hsu et al., 2020)	FCOS	ResNet-101	51.2 / 45.0
SAPNet (Li et al., 2020a)	F-RCNN	VGG-16	44.9 / 43.4
MGADA (Zhou et al., 2022)	F-RCNN	VGG-16	49.8 / 45.2
MGADA (Zhou et al., 2022)	FCOS	VGG-16	54.6 / 48.5
SIGMA (Li et al., 2022a)	F-RCNN	VGG-16	53.7 / 45.8
TDD (He et al., 2022)	F-RCNN	VGG-16	53.4 / 47.4
Ours	F-RCNN	ResNet-50	**56.8 / 53.1**

4.4 Evaluating Cross-Domain Semantic Segmentation

We then extend our framework to cross-domain semantic segmentation to demonstrate the versatility of the proposed method. We combine our method with DACS (Tranheden et al., 2021) DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b). We include the recent ProCST (Ettedgui et al., 2022) for comparison, which proposed synthesizing the source-in-target image to promote the domain adaptation performance for semantic segmentation. We perform experiments for GTA5→Cityscapes adaptation. **GTA5** (Richter et al., 2016) dataset contains 24,966 synthetic images with image resolution 1920×1080 and also the pixel level semantic annotations for 19 semantic

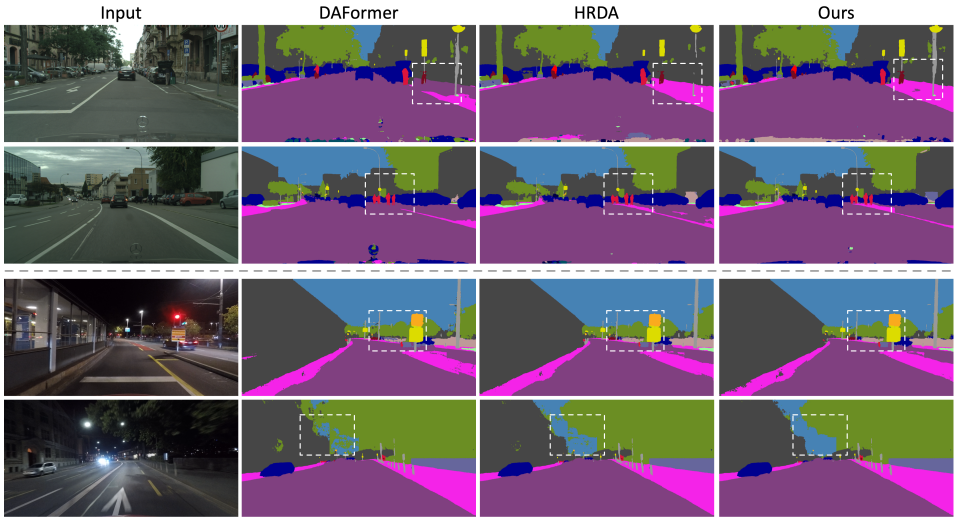

Figure 5: Qualitative comparisons with cross-domain semantic segmentation algorithms on GTA5→Cityscapes (above dashed line) and Cityscapes→Dark Zurich (below dashed line).

Table 4: Cross-domain semantic segmentation under GTA5→Cityscapes and Cityscapes→Dark Zurich-test set (**DZ** for abbreviation) adaptation.

Methods	Settings	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU ↑
DACS	GTA5→Cityscapes	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
DACS+Ours		90.4	41.2	88.4	33.1	38.2	39.2	48.1	51.5	89.4	47.2	89.1	66.9	37.4	86.0	47.3	52.1	0.2	28.1	35.1	53.1
DAFormer		95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
ProCST$_{\text{DAFormer}}$		95.4	68.2	89.8	55.1	46.4	50.4	56.4	63.4	90.4	49.9	92.3	72.4	45.3	92.6	78.4	81.2	70.6	56.8	63.6	69.4
DAFormer+Ours		93.8	56.9	90.5	58.4	43.3	58.0	63.3	71.1	91.3	48.5	94.6	74.0	16.3	94.0	83.1	85.3	72.8	64.6	68.4	69.8
HRDA		96.7	75.6	91.0	**61.0**	45.2	56.4	62.1	63.8	**91.6**	51.7	94.0	78.1	47.3	**94.1**	**84.4**	82.6	**76.6**	**63.9**	68.0	72.8
ProCST$_{\text{HRDA}}$		96.7	75.8	**91.3**	60.3	**52.8**	58.8	66.0	**72.1**	91.5	50.4	93.8	78.6	52.1	94.0	81.4	85.7	71.8	63.8	67.6	73.9
HRDA+Ours		**97.0**	**79.5**	90.9	53.2	50.2	**59.2**	**66.3**	71.3	91.5	**54.0**	**94.2**	**79.8**	**55.5**	93.8	81.4	**86.6**	75.1	62.1	**73.2**	**74.5**
DACS	Cityscapes→**DZ**	83.1	49.1	67.4	33.2	16.6	42.9	20.7	35.6	31.7	5.1	6.5	41.7	18.2	68.8	76.4	0.0	61.6	27.7	10.7	36.7
DACS+Ours		90.3	61.4	71.5	31.5	9.6	43.2	18.5	37.3	38.2	16.7	32.3	41.5	45.2	75.3	74.2	0.0	64.2	35.2	25.3	42.7
DAFormer		92.0	63.0	67.2	28.9	13.1	44.0	42.0	42.3	70.7	28.2	83.6	51.1	39.1	76.4	31.7	0.0	78.3	43.9	26.5	48.5
DAFormer+Ours		94.7	72.9	69.8	38.3	15.6	54.2	**54.8**	47.9	61.3	**39.2**	61.2	59.7	53.7	82.6	40.3	0.2	86.8	54.9	28.9	53.5
HRDA		90.4	56.3	72.0	39.5	19.5	**57.8**	52.7	43.1	59.3	29.1	70.5	60.0	**58.6**	**84.0**	**75.5**	**11.2**	**90.5**	51.6	**40.9**	55.9
HRDA+Ours		**94.8**	**73.4**	**75.3**	**40.4**	**21.8**	55.5	54.3	**48.7**	**68.7**	38.7	**78.3**	**62.2**	54.7	83.1	44.5	0.2	87.1	**57.3**	33.4	**56.4**

classes. The experimental results based on different algorithms are reported in Table 4. The proposed method could achieve various degrees of performance improvement based on different backbones.

Furthermore, we perform the Cityscapes→Dark Zurich (daytime-to-nighttime) adaptation. **Dark Zurich** dataset (Sakaridis et al., 2019) is captured in Zurich, with 3,041 daytime, 2,920 twilight and 2,416 nighttime images for training, which are all unlabeled. The Dark Zurich contains 201 annotated nighttime images: 151 images (Dark Zurich-test) are used for testing and 50 images are used for validation. The experimental results are also reported in Table 4. With a larger distribution shift (complicated mixed style and illumination factors), the proposed method could achieve more performance gains since our method can effectively reduce the visibility gap. Finally, we provide the qualitative comparisons with DAFormer and HRDA for GTA5→Cityscapes and Cityscapes→Dark Zurich adaptation in Figure 5. We conduct Synthia→Cityscapes adaptation in our appendix.

4.4.1 Evaluating Cross-Domain Depth Estimation

To evaluate depth estimation, we conducted the domain adaption from the adverse to normal condition, e.g., foggy→daytime for visibility enhancement. We evaluate the performance of Monodepth2, a recent monocular depth estimator (Godard et al., 2019), on the KITTI dataset. We use the pre-trained model on the KITTI dataset with the model resolution of 1024×320 for evaluation. In Table 5, we provide quantitative comparisons on results without and with visibility enhancement on the Foggy Cityscapes dataset since ground truth depth is provided in this dataset. As can be seen, with visibility enhancement by our domain adaptation, the depth estimator performs better than the baseline.

4.5 Ablation Studies

Effectiveness of our contrastive learning. The image synthesis is important for domain-invariant feature extraction. We choose SimCLR for comparison on the Cityscapes→Foggy Cityscapes adaptation. We compute the average LPIPS score (Zhang et al., 2018) (lower is better) between 500

Table 5: Cross-domain depth estimation. The pre-trained depth estimator achieves higher performance on Foggy Cityscapes with our foggy→daytime domain adaptation.

Method	Error↓				Accuracy↑		
	RMSE	RMSE(log)	Abs Rel	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
w/o foggy→daytime	13.74	0.430	0.319	4.587	0.445	0.737	0.875
w/ foggy→daytime	**10.70**	**0.317**	**0.232**	**2.769**	**0.578**	**0.857**	**0.947**

Table 6: Effectiveness of our reference-guided image synthesis. Compared to other image synthesis methods, our method achieves better performance as shown by the object detection task under Cityscapes→Foggy Cityscapes adaptation.

Methods	Source only	CycleGAN (Zhu et al., 2017)	TSIT (Jiang et al., 2020)	Ours
mAP ↑	31.7	36.2	50.6	**53.6**

Table 7: Effectiveness of our cross-domain content alignment (CDCA) on object detection on Foggy Cityscapes. CDCA can be added to existing methods, improving overall mAP.

Methods	CDCA	Backbone	person	rider	car	truck	bus	train	motor	bicycle	mAP↑
DA-faster (Chen et al., 2018)	×	VGG-16	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.3	27.6
	✓	VGG-16	27.4	32.8	41.7	23.5	37.4	21.4	21.5	29.1	29.4 (**+1.8**)
SCL (Shen et al., 2019)	×	VGG-16	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
	✓	VGG-16	32.5	44.9	45.6	31.5	43.1	41.8	34.8	37.1	38.9 (**+1.0**)
UMT (Deng et al., 2021)	×	VGG-16	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
	✓	VGG-16	56.6	39.1	49.5	31.5	34.2	47.3	47.3	35.0	42.6 (**+0.9**)
MeGA-CDA (Vs et al., 2021)	×	VGG-16	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
	✓	VGG-16	38.7	49.8	53.1	27.1	49.9	47.7	35.5	38.9	42.6 (**+0.8**)

validation images from Cityscapes dataset (*source*) and the according retrieved top-1 foggy images from the Foggy Cityscapes dataset (*target*) for content similarity measurement. The LPIPS score of our method is **0.1718** while the naive SimCLR achieves 0.3936. The average LPIPS score of random sampling is *0.6123*. The proposed method can effectively reduce the content mismatch between the source and target images. The t-SNE visualization is also included in our appendix.

Effectiveness of our reference-guided image synthesis in cross-domain tasks. We have reported the FID scores of different image synthesis algorithms in Table 1, which can only provide the image synthesis quality comparison. To perform further analysis on the effect of the image synthesis for perception tasks, we adopt the generated images for the object detection task under the Cityscapes→Foggy Cityscapes adaptation (Table 6). As can be seen, the proposed method could achieve the largest performance improvement compared with other algorithms. Please refer to Table 11 in our appendix for detailed per-class detection results.

Effectiveness of cross-domain content alignment (CDCA) in cross-domain tasks. We evaluate the effectiveness of the retrieved target images provided by CDCA versus traditional methods (*e.g.*, random sampling) in a downstream task. Here we choose cross-domain object detection. For CDCA, only in this experiment, to guarantee the sample diversity and that most target images will be sampled for training, instead of top-1 retrieval, we use the top-10 retrieved target images for each source image during training: we randomly select one over the 10 target images for the source image to perform adaptation in each iteration. As reported in Table 7, our proposed CDCA module can alleviate the content mismatches between the source images and the target images, achieving performance gain, and potentially working as a plug-and-play module for existing cross-domain perception algorithms.

5 Conclusion

In this paper, we comprehensively performed the analysis of content mismatch during domain adaptation. Built on a mutual-beneficial system of reference-guided image synthesis and contrastive learning, we can alleviate the content mismatch and perform the task-agnostic image synthesis for various visual perception tasks in autonomous driving. The proposed method demonstrates the effectiveness of the disentanglement of domain-invariant and domain-specific representations. The comprehensive experiments of using different benchmark algorithms on various datasets have demonstrated the superior performance of the proposed method.

REFERENCES

David Bruggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. *WACV*, 2023.

Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11457–11466, 2019.

Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7181–7190, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, volume 1, pp. 1597–1607, 2020.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348, 2018.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4091–4101, 2021.

Shahaf Ettedgui, Shady Abu-Hussein, and Raja Giryes. Procst: Boosting semantic segmentation using progressive cyclic style-transfer. *arXiv preprint arXiv:2204.11891*, 2022.

RCNN Faster. Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 9199(10.5555):2969239–2969250, 2015.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of The ACM (CACM)*, 63(11):187–208, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9580, 2022.

Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 6668–6677, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. Pmlr, 2018.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9924–9935, 2022a.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. *arXiv preprint arXiv:2204.13132*, 2022b.

Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pp. 733–748. Springer, 2020.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, pp. 172–189, 2018.

Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *European Conference on Computer Vision*, pp. 206–222. Springer, 2020.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 480–490, 2019.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, pp. 35–51, 2018.

Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 481–497. Springer, 2020a.

Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European conference on computer vision*, pp. 440–456. Springer, 2020b.

Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 1949–1957, 2021.

Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5291–5300, 2022a.

Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1347, 2021.

Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Comogan: continuous model-guided image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14288–14298, 2021.

Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8558–8567, 2021.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pp. 102–118. Springer International Publishing, 2016.

Eduardo Romera, Luis M Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1312–1318. IEEE, 2019.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 780–790, 2019.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7374–7383, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10765–10775, 2021.

Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019.

Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *European Conference on Computer Vision*, pp. 403–419. Springer, 2020.

Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9133–9142, 2021.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1379–1389, 2021.

Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, and Di Huang. Target-relevant knowledge preservation for multi-source domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5301–5310, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11724–11733, 2020a.

Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12355–12364, 2020b.

Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3273–3282, 2021.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, and Xiang Chen. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 382–391, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12425–12434, 2021.

Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pp. 86–102. Springer, 2020a.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020b.

Haitian Zheng, Haofu Liao, Lele Chen, Wei Xiong, Tianlang Chen, and Jiebo Luo. Example-guided image synthesis using masked spatial-channel attention and self-supervision. In *European Conference on Computer Vision*, pp. 422–439. Springer, 2020a.

Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *European conference on computer vision*, pp. 155–170. Springer, 2020b.

Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *arXiv preprint arXiv:2108.03553*, 2021.

Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9581–9590, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pp. 2223–2232, 2017.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

In this appendix, we provide additional implementation details (Section A), additional comparison results, more ablation studies and discussions about our method (Section B).

A Implementation Details

A.1 Contrastive Learning

We adopt SimCLR (Chen et al., 2020) as our backbone for the contrastive learning. The projected output z is a 128-dimensional vector while the temperature τ inherited from Chen et al. (2020) is 0.5. The batch size is set to 64. The image resolution is 256×256 for contrastive learning. We change the random resize scale from $(0.2, 1.0)$ adopted in SimCLR to $(1.0, 1.12)$ to preserve the most content representations in the figures. We choose Adam optimizer (Kingma & Ba, 2015) with a learning rate of $1\mathrm{e}{-3}$ and weight decay of $1\mathrm{e}{-6}$ for optimization.

A.2 Reference-guided Image Synthesis

We adopt TSIT (Jiang et al., 2020) as our backbone for reference-guided image synthesis. It is worthy to note that we adopt the instance normalization (FAdaIN) with a small batch size (2 for image resolution 1024×512 and 1 for image resolution 2048×1024) for the translation module, and batch normalization with a larger batch size (64) for the contrastive learning. For generating the images with image resolution 2048×1024, the channel number of the first convolutional module for downsampling is set to 32 for reducing the computational cost and memory burden. The multi-scale discriminator architecture remains the same for the above two settings. At the inference stage, we generate the target-like images in the target domain based on the constructed source-reference pairs generated by the cross-domain content alignment module. The image synthesis is then utilized to reduce the domain gap. Note that the translation module and the following object detection, semantic segmentation and depth estimation module are optimized separately.

A.3 Cross-Domain Object Detection

The training epoch is set to 12 for Cityscapes→Foggy Cityscapes adaptation, and 3 for both KITTI→Cityscapes and Sim10k→Cityscapes. For the multi-source cross-domain object detection setting, the training epoch is set to 9 for Cityscapes+KITTI→BDD100K and 12 for Daytime+Nighttime→Dawn on the BDD100K dataset. All the experimental results are performed based on the MMDetection following the configuration of fast_rcnn_r50_fpn_1x. Both the original source images and translated images are utilized for training the detection network. The annotations of the original source images are inherited by the translated counterparts in the target domain. The image resolution for the generated images is 1024×512 and we resize the generated images into the required image resolution (shorted side is 600 pixels) while keeping the aspect ratio unchanged.

A.4 Cross-Domain Semantic Segmentation

To preserve as much information as possible, the image resolution is set to 2048×1024 for generating images for domain adaptive semantic segmentation. We adopted DACS (Tranheden et al., 2021) DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b) as the backbones to perform experiments following the official configuration files. We rerun the HRDA on our GTX 3090 for fair comparison (hardware-wise) following https://github.com/shahaf1313/ProCST/issues/2. Similarly, annotations of source images are inherited by translated images, and both source and translated images with labels are used for training. In our ablation study, we demonstrate that the proposed cross-domain content alignment (CDCA) could promote domain adaptation performance. Thus, to reduce the content mismatch problem during the domain adaptation procedure, we conduct CDCA during the domain adaptation procedure. In detail, we replace the random sampling strategy in DACS, DAFormer and HRDA with our content-aligned strategy to sample the target images. To guarantee the sample diversity and that most target images will be sampled for training, the top-10 retrieved target images for each source image will be utilized during the training procedure: we randomly select one over the 10 target images for the source image to perform adaptation in each iteration.

B ADDITIONAL RESULTS

B.1 TASK-AGNOSTIC IMAGE SYNTHESIS

We provide the qualitative results generated by our image translation module for various image synthesis tasks: Cityscapes→Foggy Cityscapes, Sim10k/KITTI→Cityscapes, Cityscapes/KITTI→BDD100K, daytime→dawn on BDD100K, Synthia/GTA5→Cityscapes and Cityscapes→Dark Zurich in Figure 6. As illustrated, the proposed method could achieve reasonable and natural image synthesis.

Figure 6: The qualitative results of various image synthesis tasks.

B.2 CROSS-DOMAIN OBJECT DETECTION

In this section, we provide the experimental results for the **multi-source domain adaptive object detection**. We first perform the KITTI+Cityscapes→BDD100K adaptation. The images from both KITTI and Cityscapes datasets are used for training and the daytime training images from the BDD100K dataset without annotations are used for training following the experimental setup of Yao et al. (2021). The AP of the car category on the daytime validation images is reported. We provide quantitative comparisons with previous algorithms in Table 8. We achieve the best results on the Cityscapes+KITTI→BDD100K domain adaptation.

Furthermore, we perform the Daytime+Nighttime→Dawn adaptation on the BDD100K dataset (Yu et al., 2020). The daytime and nighttime images from the BDD100K dataset are used for training and the dawn images are regarded from the target domain without annotations. Similarly, following the experimental setup of Yao et al. (2021), we report the detailed experimental results under various settings in Table 9. The AP of 10 categories is reported on the dawn validation images.

B.3 CROSS-DOMAIN SEMANTIC SEGMENTATION

Synthia dataset (Ros et al., 2016) is a synthetic dataset that consists of 9,400 images with image resolution 1280×960 rendered from a virtual city and comes with pixel-level semantic annotations for 16 classes. In Table 10, we provide the comparison results between our method and previous algorithms in detail. As illustrated, the proposed method could achieve various degrees of performance gain based on different backbones.

B.4 ENHANCEMENT FOR DEPTH ESTIMATION

In this section, we provide the qualitative results of foggy→daytime translation on the Foggy Cityscapes dataset and Nighttime→Daytime translation on the BDD100K dataset and the corresponding mono depth estimation results in Figure 7. For the image enhancement, the image resolution for the reference-guided image synthesis is set to 1024×512. Then we adopt the pre-trained Monodepth2 model on the KITTI dataset with the model resolution of 1024×320 for evaluation. The qualitative results demonstrate that reasonable and consistent image enhancement could promote depth estimation performance.

Table 8: Experimental results on cross-domain object detection from Cityscapes and KITTI to BDD100k (daytime) (Yu et al., 2020). Average precision (AP, %) on car category in the target domain is evaluated. The best result is in bold.

Standards	Methods	AP on car ↑
Source-only	Cityscapes	44.6
	KITTI	28.6
	Cityscapes+KITTI	43.2
Cityscapes-only DA	Strong-Weak Saito et al. (2019)	45.5
	SCL Shen et al. (2019)	46.3
	DA-ICR-CCR Xu et al. (2020a)	45.3
	SW-ICR-CCR Xu et al. (2020a)	46.5
KITTI-only DA	Strong-Weak Saito et al. (2019)	29.6
	SCL Shen et al. (2019)	31.1
	DA-ICR-CCR Xu et al. (2020a)	29.2
	SW-ICR-CCR Xu et al. (2020a)	30.8
Source-combined DA	Strong-Weak Saito et al. (2019)	41.9
	SCL Shen et al. (2019)	43.0
	DA-ICR-CCR Xu et al. (2020a)	41.3
	SW-ICR-CCR Xu et al. (2020a)	43.6
Multi-source DA	MDAN Zhao et al. (2018)	43.2
	M^3SDA Peng et al. (2019)	44.1
	DMSN Yao et al. (2021)	49.2
	TKPD Wu et al. (2022)	58.4
	Ours†	63.1
	Ours	**66.4**
Oracle	Faster R-CNN Faster (2015)	60.2

Table 9: Experimental results on cross-domain object detection from Daytime and Nighttime to Dawn on BDD100K dataset Yu et al. (2020). Average precision (AP, %) on 10 categories in the target domain is evaluated. The best result is in bold.

Standards	Methods	bike	bus	car	motor	person	rider	light	sign	train	truck	mAP↑
Source-only	Daytime	35.1	51.7	52.6	9.9	31.9	17.8	21.6	36.3	0	47.1	30.4
	Night	27.9	32.5	49.4	15.0	28.7	21.8	14.0	30.5	0	30.7	25.0
	Daytime+Night	31.5	46.9	52.9	8.4	29.5	21.6	21.7	34.3	0	42.2	28.9
Daytime-only DA	Strong-Weak Saito et al. (2019)	34.9	51.2	52.7	15.1	32.8	23.6	21.6	35.6	0	47.1	31.4
	SCL Shen et al. (2019)	29.1	51.3	52.8	17.2	32.0	19.1	21.8	36.3	0	47.2	30.7
	GPA Xu et al. (2020b)	36.6	52.1	53.1	15.6	33.0	23.0	21.7	35.4	0	48.0	31.8
	DA-ICR-CCR Xu et al. (2020a)	35.6	47.5	52.7	13.9	32.2	22.7	22.8	35.5	0	45.7	30.9
	SW-ICR-CCR Xu et al. (2020a)	32.8	51.4	53.0	15.4	32.5	22.3	21.2	35.4	0	47.9	31.2
Night-only DA	Strong-Weak Saito et al. (2019)	31.4	38.2	51.0	9.9	29.5	22.2	18.7	32.5	0	35.7	26.9
	SCL Shen et al. (2019)	25.3	31.7	49.3	8.9	25.8	21.2	15.0	28.6	0	26.2	23.2
	GPA Xu et al. (2020b)	32.7	38.3	51.8	14.1	29.0	21.5	17.1	31.1	0	40.0	27.6
	DA-ICR-CCR Xu et al. (2020a)	30.0	32.4	50.1	14.4	29.1	22.8	17.4	32.2	0	29.7	25.8
	SW-ICR-CCR Xu et al. (2020a)	32.3	45.1	51.6	7.2	29.2	24.9	19.9	33.0	0	41.1	28.4
Source-combined DA	Strong-Weak Saito et al. (2019)	29.7	50.0	52.9	11.0	31.4	21.1	23.3	35.1	0	44.9	29.9
	SCL Shen et al. (2019)	33.9	47.8	52.5	14.0	31.4	23.8	22.3	35.4	0	45.1	30.9
	GPA Xu et al. (2020b)	31.7	48.8	53.9	20.8	32.0	21.6	20.5	33.7	0	43.1	30.6
	DA-ICR-CCR Xu et al. (2020a)	28.2	47.6	51.6	17.6	28.8	21.9	17.4	33.2	0	45.8	29.2
	SW-ICR-CCR Xu et al. (2020a)	25.3	51.3	52.1	17.0	33.4	18.9	20.7	34.8	0	47.9	30.2
Multi-source DA	MDAN Zhao et al. (2018)	**37.1**	29.9	52.8	15.8	35.1	21.6	24.7	38.8	0	20.1	27.6
	M^3SDA Peng et al. (2019)	36.9	25.9	51.9	15.1	35.7	20.5	24.7	38.1	0	15.9	26.5
	DMSN Yao et al. (2021)	36.5	54.3	55.5	20.4	36.9	**27.7**	**26.4**	**41.6**	0	**50.8**	35.0
	TKPD Wu et al. (2022)	–	–	–	–	–	–	–	–	–	–	**39.8**
	Ours	36.9	**58.8**	**68.5**	**23.4**	**39.4**	21.9	23.4	35.9	0	50.0	39.1
Oracle	Faster R-CNN Faster (2015)	27.2	39.6	51.9	12.7	29.0	15.2	20.0	33.1	0	37.5	26.6

Table 10: Comparison with UDA methods on semantic segmentation under the Synthia→Cityscapes. The best result is in bold.

Methods	Settings	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU ↑
DACS	Synthia→Cityscapes	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
DACS+Ours		83.4	26.5	82.3	24.3	4.5	39.1	24.2	25.5	85.0	–	92.4	68.4	35.4	83.5	–	39.1	–	34.3	49.6	49.8
DAFormer		84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
$\text{ProCST}_{\text{DAFormer}}$		84.8	39.5	88.2	40.2	7.3	51.1	56.3	55.1	86.5	–	89.8	74.6	48.4	86.5	–	58.6	–	55.6	62.9	61.6
DAFormer+Ours		84.1	44.2	87.8	38.7	6.2	54.9	64.3	53.7	88.1	–	93.4	78.7	48.2	89.2	–	33.1	–	63.0	64.7	62.0
HRDA		85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	–	92.9	**79.4**	**52.8**	**89.0**	–	64.7	–	63.9	64.9	65.8
$\text{ProCST}_{\text{HRDA}}$		86.7	50.7	88.7	50.0	**6.4**	**58.0**	**66.6**	**62.2**	86.4	–	**93.6**	78.8	52.7	88.4	–	64.6	–	63.5	64.4	**66.4**
HRDA+Ours		**86.9**	**51.3**	**88.9**	**50.6**	4.8	55.7	64.2	62.0	87.7	–	93.2	72.2	52.7	87.7	–	**67.2**	–	**64.5**	**66.2**	66.0

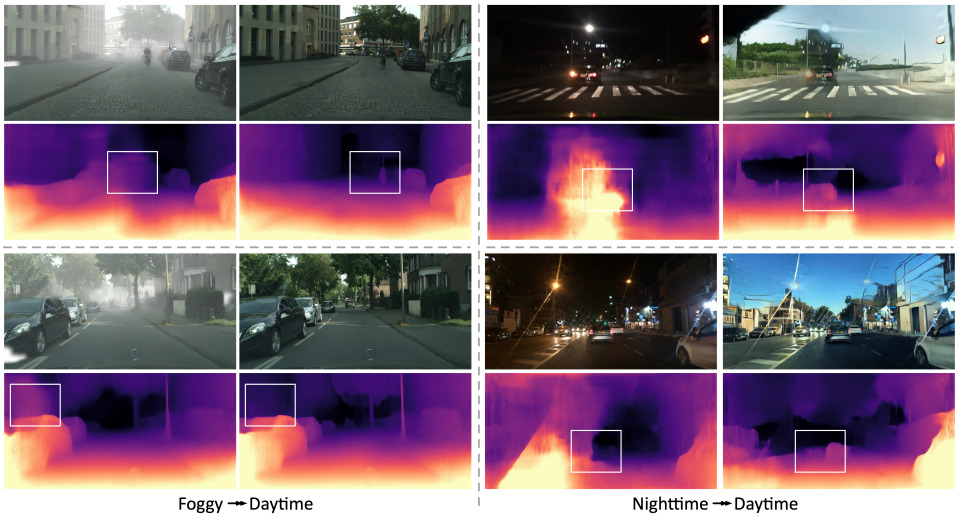

Figure 7: The qualitative results of the depth estimation based on the pre-trained model.

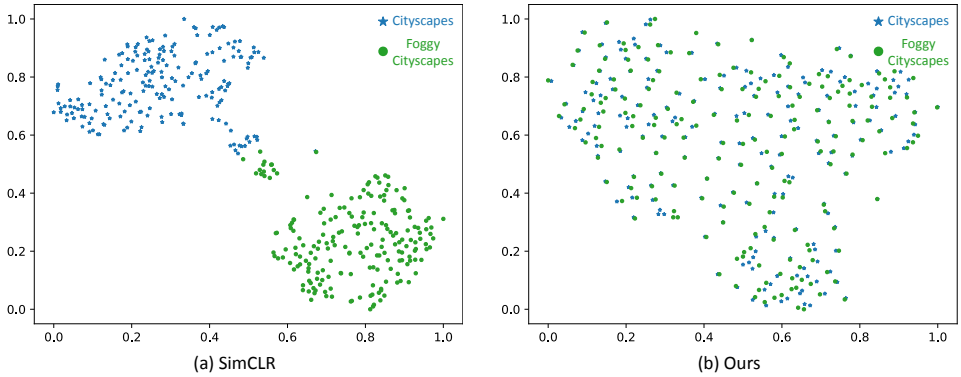

Figure 8: The t-SNE visualization of feature representation extracted from 400 samples from Cityscapes and Foggy Cityscapes datasets.

B.5 Ablation Studies

T-SNE visualization. To demonstrate that the reference-guided image synthesis is necessarily important for extracting the domain-invariant feature extraction, we compare SimCLR with our method. The naive SimCLR is optimized based on the real source images from the Cityscapes dataset and target images from the Foggy Cityscapes dataset. Differently, the proposed method with image synthesis is optimized by the source images and the corresponding translated images. We provide the t-SNE visualization comparison between SimCLR and our method as shown in Figure 8. The obtained feature representations extracted by $f(\cdot)$ for 200 clear images from the Cityscapes dataset and 200 foggy images from the Foggy Cityscapes dataset are used for visualization. We can observe that the proposed method could extract the domain-invariant content representations (more mixed distribution as shown in Figure 8).

Effectiveness of Cross-domain Content Alignment (CDCA). We provide the visual quality results of performing the cross-domain content alignment (CDCA) on Synthia→Cityscapes and GTA5→Cityscapes in Figure 9. The proposed method could return the images with similar content (layout) representation to the query image. Besides, we also provide the image synthesis results for GTA5→Cityscapes under two settings: without CDCA and with CDCA in Figure 10.

Effectiveness of image synthesis for cross-domain object detection. Additionally, we provide the detailed per-class detection results in Table 11 through using the translated images by different image synthesis algorithms for Cityscapes→Foggy Cityscapes adaptation. In theory, better image synthesis

Figure 9: The cross-domain content alignment (image retrieval) results for Synthia→Cityscapes (above the gray dashed line) and GTA5→Cityscapes (below the gray dashed line). The first column is the query image and the images in the other columns are the top-4 retrieved images.

Table 11: The per-class results of comparison on cross-domain object detection on the Foggy Cityscapes dataset for Cityscapes→Foggy Cityscapes adaptation. The best result is in bold.

Methods	person	rider	car	truck	bus	train	motor	bicycle	mAP↑
Source only	40.7	46.1	45.0	19.5	27.9	3.6	27.4	43.6	31.7
CycleGAN (Zhu et al., 2017)	44.3	52.0	50.3	25.3	29.6	9.5	32.1	46.6	36.2
TSIT (Jiang et al., 2020)	**54.1**	**58.5**	**72.8**	34.5	54.5	36.1	41.5	**53.1**	50.6
Ours	53.7	58.3	72.2	**36.6**	**60.6**	**51.3**	**44.3**	51.6	**53.6**

will lead to better object detection performance in the target domain. As reported, the proposed method has achieved the largest performance gain over all the image synthesis algorithms.

B.6 LIMITATION

While the proposed method aims to reduce the appearance shift between the source domain and the target domain by image synthesis in the target domain, it shows limited ability on generating abstract, highly iconic, exaggerated, or succinct images (*e.g.*, photo-to-caricature, photo-to-watercolor), resulting in marginal performance improvement under the unsupervised domain adaptation setting. We attribute this to the reason that the image-to-image translation module generates unfaithful target-like images due to the change of scale, shape, and other image characteristics between the source and target domain. Incorporating more sophisticated image synthesis methods might resolve this problem in future work.

Figure 10: The image synthesis results of without CDCA (above the dashed line) and with CDCA (below the dashed line) for GTA5→Cityscapes adaptation. Best viewed in color.