Cross-Domain Autonomous Driving Perception using Contrastive Appearance Adaptation

Ziqiang Zheng[1], Yingshu Chen[1], Binh-Son Hua[2,3], Yang Wu[4], Sai-Kit Yeung[1]

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 domains, thus 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 source-target image pairs. Our framework consists of three mutually 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* based on the generated content-aligned source-target image pairs, and a *contrastive learning* module to self-supervise domain-invariant feature extractor. 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.**

I. Introduction

Building scalable and robust perception capabilities such as object detection [1], [2], semantic segmentation [3] and depth estimation [4] is a challenging task in autonomous driving systems [5], [6], [7]. A fundamental challenge is the domain shift, where the systems are expected to work in various conditions such as adverse weather, changing illumination, and varying geographic locations. In theory, supervised learning can be used to train object detectors, semantic segmentation models, and depth estimators with paired data and labels acquired in various conditions, but in practice, the supervised learning 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 focus mainly on feature-level adaptation, including discrepancy-based [8], adversarial feature learning [1], self-training [9] and knowledge distillation [10]. Beyond these methods, image-level adaptation [11] 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 cross-domain perception 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, which degrades overall domain adaptation performance.

In this paper, we observe 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 limited semantic "objectness" correspondences or layout (geometry) similarity [12] between a source-domain image and a target-domain image. We provide a motivating example in Figure 1. In this example, our aim is 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 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 us to choose 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 domain-invariant and domain-specific representations.

To better model these semantic correspondences, we propose a novel framework for image-level domain adaptation that incorporates cross-domain content alignment (CDCA), contrastive learning, and reference-guided image synthesis. Our method is general-purpose and task-agnostic and can support different perception tasks such as object detection, semantic segmentation, and depth estimation. The modules in our framework are also mutually beneficial. The 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 a content representation can be extracted from a domain-invariant feature extractor trained by contrastive learning

[1] Ziqiang Zheng, Yingshu Chen and Sai-Kit Yeung are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. Corresponding author: Ziqiang Zheng (zhengziqiang1@gmail.com).

[2] Binh-Son Hua is with VinAI Research, Hanoi, Vietnam.

[3] Binh-Son Hua is also with Trinity College Dublin.

[4] Yang Wu is with Tencent AI Lab, Shenzhen, China.

This research project is partially supported by an internal grant from HKUST (R9429), the Innovation and Technology Support Programme of the Innovation and Technology Fund (Ref: ITS/200/20FP), and the Marine Conservation Enhancement Fund (MCEF20107).

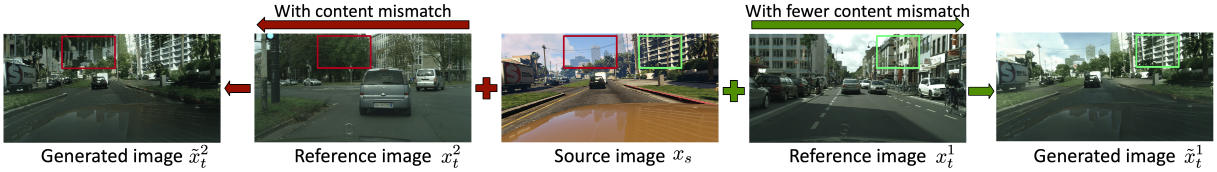
With content mismatch
With fewer content mismatch
Generated image $\tilde{x}_t^2$
Reference image x_t^2
Source image x_s
Reference image x_t^1
Generated image $\tilde{x}_t^1$

Fig. 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.

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 synthesized 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 improves overall performance. The three modules mutually improve each other and eventually converge to better domain adaptation.

We have conducted experiments with 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 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 implicit semantic correspondences between domains, resulting in better disentanglement of 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.

II. 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 (*e.g.*, discrepancy-based methods [8], adversarial training [1], self-training [9] and knowledge distillation [10]). Adversarial training was pioneered in object detection by reducing the domain discrepancy in a min-max manner with a domain classifier [1]. Similarly, one can also minimize domain discrepancy by learning domain-invariant representations for semantic segmentation [13]. However, using adversarial learning in the feature space only achieved a marginal improvement in accuracy in complex scenarios. Self-training algorithms [9] utilize the pre-trained model in the source domain to generate the supervision in the target domain for retraining. However, self-training methods suffer from the low quality of pseudo-labels generated in the target domain.

Recent knowledge distillation algorithms [14], [15] introduced the Mean Teacher framework for domain-adaptive object detection and semantic segmentation. AT [16] aimed to improve the quality of pseudo-labels generated in the target domain by using adversarial learning and mutual learning. DaFormer [10] and HRDA [17] achieved the current state-of-the-art domain adaptive semantic segmentation performance by introducing SegFormer [3] for more effective knowledge distillation. ProCST [18] introduced progressive style transfer into DAFormer and HRDA to achieve performance gain for domain adaptive semantic segmentation, which is not task-agnostic since ProCST designs the label loss based on dense pixel-level annotations. Progressive image synthesis among multiple image resolutions also requires huge computational costs and memory burdens. Although successful in some scenarios, these domain adaptation approaches remain limited when there exist significant content mismatches between the source and target images.

Content-aware adaptation. Some efforts have been made to address content mismatches between the source and target images. CCM [12] constructed positive pairs for better domain adaptation in the label space through pixel-wise similarity matching, which is not task-agnostic and requires dense pixel-level semantic annotations. Compared to CCM [12], our similarity matching is learned without annotations and guided by domain-invariant features from image-level contrastive learning and reference-guided image synthesis. Recent works [19], [20] propose to construct normal-adverse image pairs that have a similar layout based on GPS information. The utilization of the normal images collected with good visibility results in a better domain adaptation performance [21]. Refign [21] adopted a pre-trained geometry alignment module to warp the paired reference image to refine the pseudo-labels generated in the target domain. However, they require large-scale additional data for training as well as expensive geometry correspondences from structure-from-motion. Compared to these works, our work focuses on more generic cross-domain content alignment without using annotations or additional training data. We support multiple downstream tasks by building our method upon task-agnostic image synthesis.

Image synthesis for task-agnostic domain adaptation. Image synthesis methods can learn to generate target-like images to reduce the domain gap for task-agnostic image-level domain adaptation [22]. State-of-the-art image translation methods are CycleGAN with the cycle-consistency loss [11] and its variants [23], [22]. Recent work [23], [24], [24] introduced image translations for cross-domain image classification [24], semantic segmentation [23] and object detection [15]. 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 [25] 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. However, these methods do not consider content mismatches between the source image and the reference image. In this work, we propose an effective approach that wires contrastive learning and image translation in a mutually beneficial way for retrieving content-aware reference images, thereby reducing content mismatches and improving overall image translation and domain adaptation performance.

III. Our Method

A. Overview

We first formulate our problem by assuming the domain adaptation from a source domain $\mathbb{S}$ to a target domain $\mathbb{T}$, where their data distributions are different, i.e., $\mathcal{P}_{\mathbb{S}} \neq \mathcal{P}_{\mathbb{T}}$. The source domain 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 (consistent with *source-target* pairs in this paper); 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 target-like 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, building a mutual-beneficial training cycle.

B. Cross-domain Content Alignment

In each cycle, to perform cross-domain content alignment, all real source and target images from the whole training datasets are fed into the feature extractor $f(\cdot)$ to obtain the content representations. $f(\cdot)$ is a ResNet-50 [26] and is initialized with the pre-trained weights on ImageNet dataset for better initialization performance and optimized by contrastive learning described in Section III-D. To construct 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.

C. Reference-guided Image Synthesis

After performing CDCA, the constructed pairs are then used for reference-guided image synthesis, generating target-like images to reduce the domain gap. Let the source-reference image pair generated by CDCA be (x_s, x_t). The 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 output. x_s is fed into the *content stream* to extract the feature of the source content, while x_t is fed into the *style stream* to extract domain-specific representations (style or appearance). We use FAdaIN and FADE [25] to transfer the style representation from x_t and preserve the content representation of x_s. 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. 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 [27] to enforce the similarity of the intermediate feature representations at different layers of the multi-scale discriminators.

D. Domain-invariant Representation Learning

The contrastive learning framework is adopted to project both the source and target images into the same feature space. Particularly, we 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 use contrastive learning to train the feature extractor to be insensitive to the domain gap between x_s and $\tilde{x}_t$. In other words, we adopt image synthesis as an effective data augmentation for contrastive learning. We feed the source image x_s and the 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 $z_{\tilde{x}_t}$ to calculate the contrastive loss. We also apply transform operations from the transformation sets

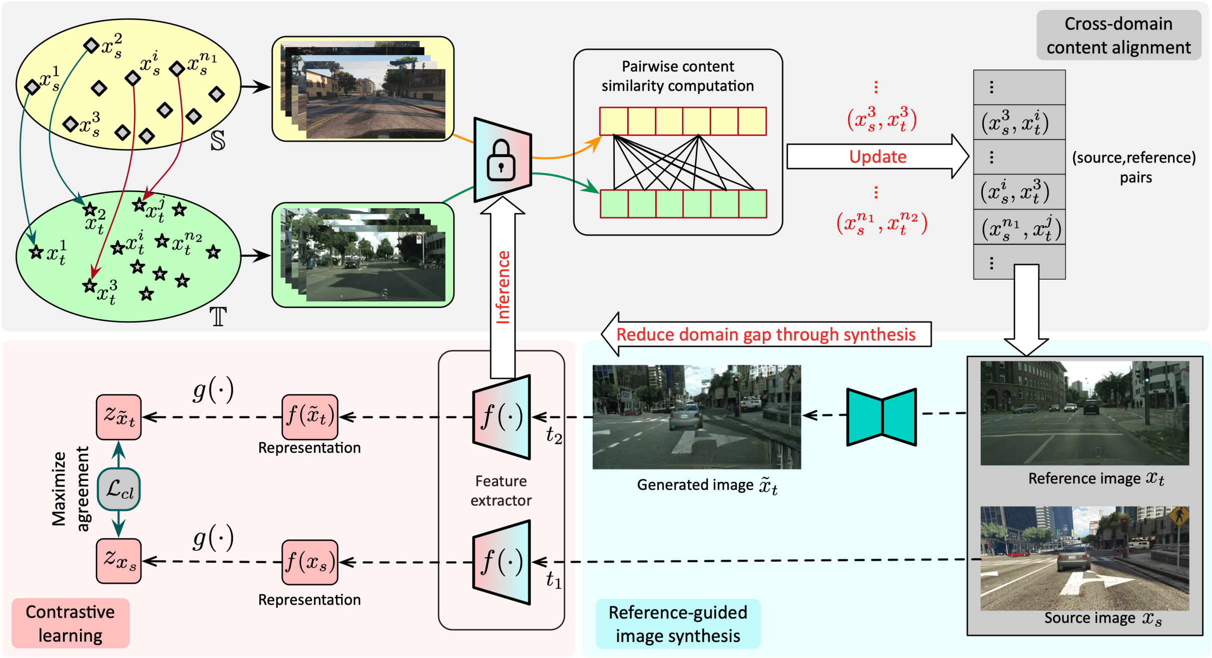

Fig. 2. The core of our image-level domain adaptation is a cycle of three mutual-beneficial modules: cross-domain content alignment (CDCA), reference-guided image synthesis and contrastive learning. CDCA uses the domain-invariant feature extractor learned by contrastive learning to construct source-reference 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.

$\mathcal{T}$ (including *random resizing and cropping*, *color jitter* and *random greyscale*) to obtain $x_s \leftarrow t_1(x_s)$ and $\tilde{x}_t \leftarrow t_2(\tilde{x}_t)$ where t_1, t_2 are augmentation operators randomly drawn from $\mathcal{T}$. The contrastive loss $\mathcal{L}_{cl}$ is written as

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

where $sim(u, v) = u^T v / \|u\| \|v\|$ denotes the pairwise 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.

E. Cross-domain Perception

We train the modules in our framework in an end-to-end manner. Particularly, we integrate reference-guided image synthesis and contrastive learning through joint training and iteratively optimize each. To alleviate error propagation, we do not allow gradient backpropagation from the contrastive learning module to the synthesis module. The synthesis module and contrastive learning module are optimized separately. Making such a framework work seamlessly with good performance is creative and non-trivial. Each module in our system can also be replaced with counterparts and integrating our CDCA module with other frameworks as a plug-and-play module could also achieve performance gain. The training of the proposed CDCA module does not exhibit instability and the whole framework is optimized steadily. We include more analysis and the training curves in our ablation studies. 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 both the source images and the generated target-like images with the same source labels for supervised learning. For depth estimation, we consider unsupervised monocular depth estimation and assume that a well-trained depth estimator is available in the source domain. We then perform the target→source domain adaptation and apply the depth estimator to the generated source-like images.

IV. Experiments

A. Implementation Details

For the reference-guided image synthesis module, we adopt TSIT [25] as the backbone and remove the perceptual loss part. We adopt a small batch size for image synthesis: 2 for image resolution 1024×512 (for cross-domain object detection and depth estimation tasks) and 1 for image resolution 2048×1024 (for cross-domain semantic segmentation). For the contrastive learning module, we adopt ResNet-50 [26] pre-trained on ImageNet for $f(\cdot)$ and preserve the same architecture for $g(\cdot)$ as [28]. The projected output z is a 128-dimensional vector while the temperature τ is 0.5. The batch size is set to 64 and 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 with a learning rate of $1e{-}3$ and weight decay of $1e{-}6$ for contrastive learning optimization. Besides, to iteratively optimize the modules, in each cycle, the epoch

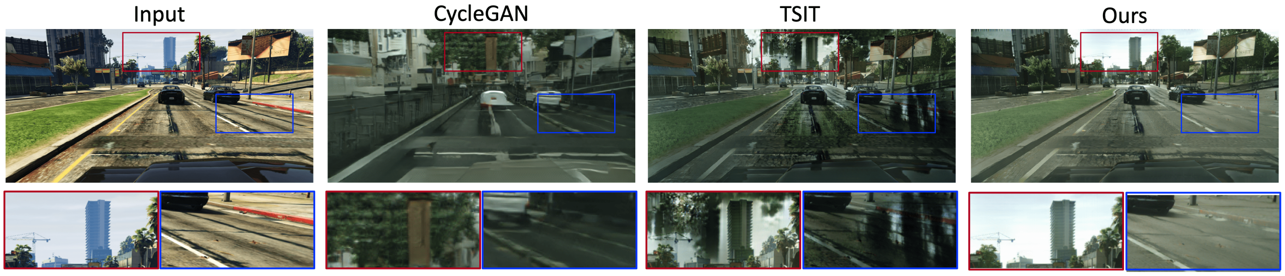

Fig. 3. Generated image quality comparison between CycleGAN, TSIT and our method under Sim10k→Cityscapes adaptation.

TABLE I

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

Methods	Cityscapes→Foggy Cityscapes	Sim10k→Cityscapes
CycleGAN [11]	25.76	77.31
TSIT [25]	7.35	60.24
Ours	**5.68**	**48.23**

number of the synthesis module and contrastive learning module is set to 5 and 10 and we execute the cycle 10 times in total. We perform cross-domain object detection experiments based on the MMDetection framework [29]. We adopt Faster R-CNN [2] with VGG-16 [30] and ResNet-50 [26] pre-trained on ImageNet as the backbone network. The shorter side of each input image is resized to 600 pixels. For semantic segmentation, we adopt DACS [31] DAFormer [10] and HRDA [17] as backbones and conduct the experiments following the official instructions. We choose both the source data and the target-like translated data to optimize both detection and segmentation models, which encourages the training model without being biased.

B. Task-agnostic Image Synthesis

We first evaluate the quality of the generated images qualitatively and quantitatively as high-quality image synthesis could imply high performance for downstream perception tasks. We compare our reference-guided image synthesis with two representative image synthesis methods namely CycleGAN [11] and TSIT [25]. CycleGAN translates the source image to the target domain without reference. TSIT is a reference-guided image synthesis method. For quantitative measurement of the synthesized image quality, we adopt FID [32] (lower is better) as the evaluation metric. We report qualitative and quantitative results in Figure 3 and Table I 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.

C. Cross-Domain Perception

Cross-Domain Object Detection. We first report quantitative results on Cityscapes→Foggy Cityscapes in Table II. The average precision (AP) of 8 categories on the Foggy Cityscapes and the mAP are reported. We compare our method with recent state-of-the-art algorithms including SCL [33], GPA [34], UMT [15], MeGA-CDA [35], CDG [36], MGADA [37], SIGMA [38], TDD [39] and AT [16]. Our method outperforms existing methods by a large margin even based on the same backbone. Qualitative results are shown in Figure 4. As illustrated, our method could accurately detect small objects in dense fog, *e.g.*, the bicycle in the first row. We also conduct **synthetic-to-real** and **cross-camera** detection from Sim10k/KITTI to Cityscapes in Table III. 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. We provide more experimental results of **multi-source cross-domain** object detection in our supplementary.

TABLE II

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 [33]	F-RCNN	VGG-16	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
GPA [34]	F-RCNN	ResNet-50	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
UMT [15]	F-RCNN	VGG-16	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
MeGA-CDA [35]	F-RCNN	VGG-16	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
CDG [36]	F-RCNN	VGG-16	38.0	47.4	53.1	34.2	47.5	41.1	38.3	38.9	42.3
MGADA [37]	F-RCNN	VGG-16	43.9	49.9	60.6	29.6	50.7	39.0	38.3	42.8	44.3
SIGMA [38]	F-RCNN	ResNet-50	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
TDD [39]	F-RCNN	ResNet-50	50.7	53.7	68.2	35.1	53.0	45.1	38.9	49.1	49.2
AT [16]	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	VGG-16	53.2	**59.2**	**73.1**	35.1	56.6	40.9	42.5	50.8	51.4
Ours	F-RCNN	ResNet-50	53.7	58.3	72.2	**36.6**	**60.6**	51.3	**44.3**	51.6	**53.6**

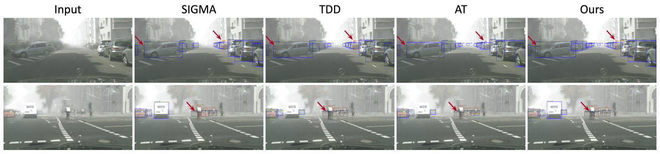

Fig. 4. Qualitative comparisons of cross-domain object detection methods on Cityscapes→Foggy Cityscapes adaptation.

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 [31], DAFormer [10] and HRDA [17]. We include the recent ProCST [18] for comparison, which proposed the synthesis of source-in-target images to improve the performance of domain adaptation for semantic segmentation. We perform Cityscapes→Dark Zurich (daytime-to-nighttime) adaptation. **Dark Zurich** dataset [19] contains 201 annotated nighttime images: 151 images (Dark Zurich-test) are used for testing and 50 images are used for validation. Following the official implementation of ProCST, we conduct the Cityscapes→Dark Zurich adaptation based on

TABLE III

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

Methods	Detector	Backbone	mAP (*car*) Sim10k / KITTI ↑
CST [40]	F-RCNN	VGG-16	44.5 / 43.6
MeGA-CDA [35]	F-RCNN	VGG-16	44.8 / 43.0
UMT [15]	F-RCNN	VGG-16	43.1 / -
CDN [41]	F-RCNN	VGG-16	49.3 / 44.9
CFA [42]	FCOS	VGG-16	49.0 / 43.2
CFA [42]	FCOS	ResNet-101	51.2 / 45.0
SAPNet [43]	F-RCNN	VGG-16	44.9 / 43.4
MGADA [37]	F-RCNN	VGG-16	49.8 / 45.2
MGADA [37]	FCOS	VGG-16	54.6 / 48.5
SIGMA [38]	F-RCNN	VGG-16	53.7 / 45.8
TDD [39]	F-RCNN	VGG-16	53.4 / 47.4
Ours	F-RCNN	VGG-16	55.3 / 50.4
Ours	F-RCNN	ResNet-50	**56.8** / **53.1**

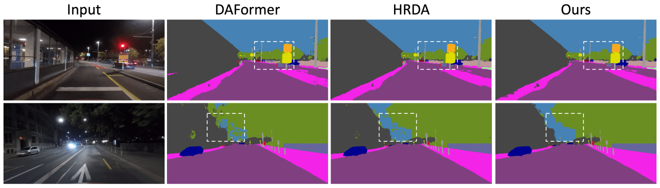

Fig. 5. Qualitative comparisons with cross-domain semantic segmentation algorithms on Cityscapes→Dark Zurich adaptation.

DAFormer and HRDA backbones. All experimental results are reported in Table IV. As reported, with a large distribution shift (complicated mixed style and illumination factors), the proposed method could achieve more performance gains than existing algorithms since our method can effectively reduce the visibility gap. Finally, we provide qualitative comparisons with DAFormer and HRDA for Cityscapes→Dark Zurich adaptation in Figure 5. We provide the results of GTA5→Cityscapes and Synthia→Cityscapes adaptation in our supplementary.

Cross-Domain Depth Estimation. To evaluate depth estimation, we perform adverse-to-normal domain adaption, *e.g.,* foggy→daytime for visibility enhancement. We evaluate the performance of Monodepth2, a recent monocular depth estimator [4], 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 V, we provide quantitative comparisons of without and with visibility enhancement on the Foggy Cityscapes dataset since the ground truth depth is provided. With visibility enhancement by our method, the depth estimator performs better than the baseline.

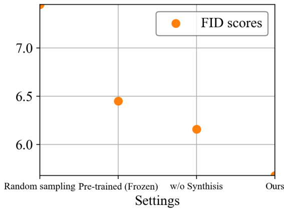

Fig. 6. Reference-guided image synthesis performance comparison based on constructed source-reference pairs under different settings.

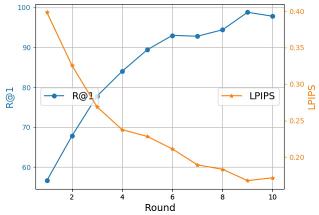

Fig. 7. R@1 and LPIPS curves computed by source-reference pairs constructed by our CDCA module during the whole training procedure.

D. Ablation Studies

Training stability. We discuss the training stability of our method. One potential issue is that at the early stages of the training, general feature extractors such as the pre-trained ResNet50 on ImageNet might yield content-mismatched source-reference pairs that lead to negative transfers, which are detrimental to model convergence. However, we empirically found such training instability does not occur in our method.

We conduct Cityscapes→Foggy Cityscapes adaptation for explanation. We explore whether the CDCA module could effectively return the reference images consistent with the given source image. The random source-reference pair construction is also conducted for comparison. We also compare features initialized from the pre-trained ResNet-50 model on ImageNet with random initialization. For quantitative results, we calculated the average LPIPS score [44] (lower is better) between 500 validation images from the Cityscapes dataset (*source*) and the retrieved top-1 images from the Foggy Cityscapes dataset (*reference*) to measure the content similarity of the source-reference pairs. Since the Foggy Cityscapes dataset is simulated from the Cityscapes dataset with one-to-one clear-foggy correspondence, we calculate the R@1 score (higher is better). As shown in Table VI, our CDCA module can effectively construct source-reference pairs even with features from a pre-trained ResNet-50 model. By further using the reference-guided image synthesis for contrastive learning, we achieve better domain-invariant feature extraction. We then adopt such constructed source-reference image pairs for reference-guided image synthesis with an FID plot provided in Figure 6 to show the image synthesis performance. “Random sampling” indicates randomly constructed source-reference pairs; “Pre-trained (Frozen)” indicates constructing source-reference pairs based on a frozen pre-trained ResNet-50 model on ImageNet; “w/o Synthesis” indicates constructing source-reference pairs under the setting where no image synthesis as augmentation for contrastive learning. We also adopt different retrieved reference images (*e.g.*, the top-1, top-5 and top-10) by our CDCA module to perform reference-guided image synthesis and report the corresponding FID scores of the generated images. The FID scores are 5.68 (consistent with “Ours” in Figure 6), 5.73 and 5.80 respectively using the top-1, top-5 and top-10 retrieved images. This analysis confirms that our method establishes plausible source-reference pairs for stable training.

Finally, the training stability of our method during the whole training procedure can also be observed in the plot of R@1/LPIPS as in Figure 7. As illustrated, the whole cycle of the proposed method could be optimized steadily and the CDCA module could progressively yield more effective source-reference pairs with consistent feature representations. Therefore, we conclude that, though CDCA might result in negative transfer, this does not happen at scale; within a large batch of images required for contrastive learning the majority of transfers are appropriate, alleviating the influence of negative transfer.

Effectiveness of CDCA in downstream tasks. We evaluate whether the formulated source-reference correspondences could boost the domain adaptation performance of downstream visual tasks (*e.g.*, domain adaptive object detection) by replacing the random sampling strategy with such cor-

TABLE IV

Cross-domain semantic segmentation under Cityscapes→Dark Zurich-test set (**DZ** for short) 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 [31]	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 [10]		93.5	65.5	73.3	39.4	19.2	53.3	44.1	44.0	59.5	34.5	66.6	53.4	52.7	82.1	52.7	9.5	89.3	50.5	38.5	53.8
ProCST$_{\text{DAFormer}}$ [18]		94.7	72.7	73.3	40.2	20.2	53.1	53.2	47.8	62.1	36.7	73.8	60.4	55.4	82.6	35.4	1.8	87.2	55.7	33.5	54.7
DAFormer+Ours		94.9	72.2	73.9	41.1	16.1	58.1	54.4	52.9	70.9	37.5	73.8	54.8	51.2	89.6	44.3	8.2	88.5	56.9	34.9	56.5
HRDA [17]		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
ProCST$_{\text{HRDA}}$ [18]		94.8	73.7	75.6	40.9	22.3	56.0	55.0	49.1	69.2	**39.3**	78.8	62.5	55.0	83.5	45.0	0.9	87.5	57.7	33.7	56.8
HRDA+Ours		**95.7**	**77.4**	**83.6**	**50.4**	**34.2**	**62.5**	**62.2**	**69.9**	**81.1**	16.7	**91.5**	**67.3**	**60.0**	**88.1**	5.5	**32.1**	**90.8**	**55.7**	**41.4**	**61.4**

TABLE V

Cross-domain depth estimation. The pre-trained depth estimator achieves higher performance on Foggy Cityscapes with foggy→daytime 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 VI

Comparison of different source-reference pairs construction strategies. Using both pretrained ResNet-50 for initialization and image synthesis for domain-invariant feature learning yields the best source-reference pairs.

Strategies		R@1 ↑	LPIPS ↓
Random source-reference pairs		0.2%	0.6123
$f(.)$ initialization	Image synthesis for contrastive learning	R@1 ↑	LPIPS ↓
Random initialization	-	14.6%	0.5314
Pretrained ResNet-50 (frozen)	-	50.4%	0.4263
Pretrained ResNet-50 (finetuned)	-	54.2%	0.3936
Random initialization	✓	89.8%	0.2034
Pretrained ResNet-50 (finetuned)	✓	**97.8%**	**0.1718**

respondences. To guarantee the sample diversity during the training procedure, instead of top-1 retrieval, we use the top-10 retrieved target images for each source image: we randomly select one target sample over the 10 target images for the source image to perform adaptation in each iteration. In Table VII, our proposed CDCA module can alleviate content mismatches between two domains, achieving performance gain, and potentially working as a plug-and-play module for existing cross-domain perception algorithms.

Effectiveness of referenced-guided image synthesis in downstream tasks. We further analyze the effectiveness of the generated images by different image synthesis algorithms for cross-domain object detection (Cityscapes→Foggy Cityscapes adaptation). For CycleGAN, TSIT and our method, both source images and generated target-like images are used for training, and the quantitative results are reported in Table VIII. The proposed method could achieve the best performance improvement compared to other algorithms. Additionally, the experimental results of only using the synthesized target-like images (denoted by 'Ours$^-$') for training are also reported, which indicates that combining both source images and target-like images can result in more performance gain.

V. Conclusion

In this paper, we comprehensively perform the analysis of content mismatches during domain adaptation. Based on a mutually beneficial system of reference-guided image synthesis and contrastive learning, our method can alleviate content mismatches and perform task-agnostic image synthesis for various visual perception tasks in autonomous driving. Comprehensive experiments using different benchmark algorithms on various datasets have demonstrated the superior performance of the proposed method.

Our method is not without limitations. While our method is effective at reducing content mismatches between domains, the proposed method did not explicitly measure objectness correspondences between domains. For further improvement, the proposed method could adopt annotations from the downstream visual tasks for domain adaptation. Additionally, exploring more downstream tasks such as 3D object detection, pedestrian detection, instance segmentation with our framework would be interesting future work.

TABLE VII

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	person	rider	car	truck	bus	train	motor	bicycle	mAP↑
DA-faster [1]	×	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.3	27.6
	✓	27.4	32.8	41.7	23.5	37.4	21.4	21.5	29.1	29.4 (**+1.8**)
SCL [33]	×	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
	✓	32.5	44.9	45.6	31.5	43.1	41.8	34.8	37.1	38.9 (**+1.0**)
UMT [15]	×	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
	✓	56.6	39.1	49.5	31.5	34.2	47.3	47.3	35.0	42.6 (**+0.9**)
MeGA-CDA [35]	×	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
	✓	38.7	49.8	53.1	27.1	49.9	47.7	35.5	38.9	42.6 (**+0.8**)

TABLE VIII

Effectiveness of reference-guided image synthesis on cross-domain object detection on the Foggy Cityscapes dataset using Cityscapes→Foggy Cityscapes adaptation. The best result is in bold. 'Ours$^-$' indicates the setting of using only the synthesized target-like images for training.

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 [11]	44.3	52.0	50.3	25.3	29.6	9.5	32.1	46.6	36.2
TSIT [25]	**54.1**	**58.5**	72.8	34.5	54.5	36.1	41.5	**53.1**	50.6
Ours$^-$	53.9	57.1	**74.4**	**36.9**	55.3	34.0	44.2	52.6	51.1
Ours	53.7	58.3	72.2	36.6	**60.6**	**51.3**	**44.3**	51.6	**53.6**

References

[1] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. 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.

[2] R. Faster, "Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 9199, no. 10.5555, pp. 2969239–2969250, 2015.

[3] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3828–3838, 2019.

[5] C. Häne, T. Sattler, and M. Pollefeys, "Obstacle detection for self-driving cars using only monocular cameras and wheel odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5101–5108, IEEE, 2015.

[6] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2267–2273, IEEE, 2020.

[7] F. C. Borlino, S. Bucci, and T. Tommasi, "Contrastive learning for cross-domain open world recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10133–10140, IEEE, 2022.

[8] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, pp. 97–105, PMLR, 2015.

[9] Y. Zou, Z. Yu, B. Kumar, and J. 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.

[10] L. Hoyer, D. Dai, and L. 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, 2022.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, pp. 2223–2232, 2017.

[12] G. Li, G. Kang, W. Liu, Y. Wei, and Y. Yang, "Content-consistent matching for domain adaptive semantic segmentation," in *European Conference on Computer Vision (ECCV)*, pp. 440–456, Springer, 2020.

[13] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. 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.

[14] Y. Zhang, Z. Wang, and Y. 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.

[15] J. Deng, W. Li, Y. Chen, and L. 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.

[16] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda, "Cross-domain adaptive teacher for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022.

[17] L. Hoyer, D. Dai, and L. Van Gool, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," *arXiv preprint arXiv:2204.13132*, 2022.

[18] S. Ettedgui, S. Abu-Hussein, and R. Giryes, "Procst: Boosting semantic segmentation using progressive cyclic style-transfer," *arXiv preprint arXiv:2204.11891*, 2022.

[19] C. Sakaridis, D. Dai, and L. V. 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.

[20] C. Sakaridis, D. Dai, and L. 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.

[21] D. Bruggemann, C. Sakaridis, P. Truong, and L. Van Gool, "Refign: Align and refine for adaptation of semantic segmentation to adverse conditions," *WACV*, 2023.

[22] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *European Conference on Computer Vision (ECCV)*, pp. 155–170, Springer, 2020.

[23] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, pp. 1989–1998, Pmlr, 2018.

[24] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8690–8699, 2021.

[25] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "Tsit: A simple and versatile framework for image-to-image translation," in *European Conference on Computer Vision*, pp. 206–222, Springer, 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. 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.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, vol. 1, pp. 1597–1607, 2020.

[29] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

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

[31] W. Tranheden, V. Olsson, J. Pinto, and L. 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.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural information processing systems (Neurips)*, vol. 30, 2017.

[33] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses," *arXiv preprint arXiv:1911.02559*, 2019.

[34] M. Xu, H. Wang, B. Ni, Q. Tian, and W. 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, 2020.

[35] V. Vs, V. Gupta, P. Oza, V. A. Sindagi, and V. 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.

[36] S. Li, J. Huang, X.-S. Hua, and L. Zhang, "Category dictionary guided unsupervised domain adaptation for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1949–1957, 2021.

[37] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. 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.

[38] W. Li, X. Liu, and Y. 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, 2022.

[39] M. He, Y. Wang, J. Wu, Y. Wang, H. Li, B. Li, W. Gan, W. Wu, and Y. 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.

[40] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *European Conference on Computer Vision*, pp. 86–102, Springer, 2020.

[41] P. Su, K. Wang, X. Zeng, S. Tang, D. Chen, D. Qiu, and X. Wang, "Adapting object detectors with conditional domain normalization," in *European Conference on Computer Vision*, pp. 403–419, Springer, 2020.

[42] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *European Conference on Computer Vision*, pp. 733–748, Springer, 2020.

[43] C. Li, D. Du, L. Zhang, L. Wen, T. Luo, Y. Wu, and P. Zhu, "Spatial attention pyramid network for unsupervised domain adaptation," in *European Conference on Computer Vision*, pp. 481–497, Springer, 2020.

[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.