Cross-Domain Autonomous Driving Perception using Contrastive Appearance Adaptation

Ziqiang Zheng\textsuperscript{1}, Yingshu Chen\textsuperscript{1}, Binh-Son Hua\textsuperscript{2,3}, Yang Wu\textsuperscript{4}, Sai-Kit Yeung\textsuperscript{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 $S$ to the target domain $T$ by using a reference image $x_1$ or $x_2$ from $T$. We observe that reference $x_1$ shares more similar semantic correspondences (green region) with the source image than reference $x_2$ (red region). Therefore, the generated image $\tilde{x}_1$ has better image quality and fewer visual artifacts than $\tilde{x}_2$. 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.
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 $S$ to a target domain $T$, where their data distributions are different, i.e., $P_S \neq P_T$. The source domain is labeled, while the target domain is unlabeled. Let $S = \{x^s_i, y^s_i\}$, 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, $T = \{x^t_i\}$ 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 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 contrastive 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 $\hat{x}_t$ is target alike, which means that the pair $(x_s, \hat{x}_t)$ 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 $L_G$ and the discriminator loss $L_D$ can be written as:

$$L_G = -\mathbb{E}[D(\hat{x}_t)] + \lambda_{fm} L_{fm}(\hat{x}_t, x_t),$$

$$L_D = -\mathbb{E}[[\min(-1 + D(x_t), 0)] - \mathbb{E}[[\min(-1 - D(\hat{x}_t), 0)],$$

where $D$ is the discriminator; $L_{fm}$ 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 $(x_s, \hat{x}_t)$ 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 $\hat{x}_t$. 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 $\hat{x}_t$ 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 $z_{\hat{x}_t}$ to calculate the contrastive loss. We also apply transform operations from the transformation sets.
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.

\[ T \text{(including random resizing and cropping, color jitter and random greyscale)} \] 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 \( T \). The contrastive loss \( L_{cl} \) is written as

\[
L_{cl} = -\log \frac{\exp(\text{sim}(z_{x_s}, z_{\tilde{x}_t})) / \tau}{\sum_{x \in X, x \neq x_s} \exp(\text{sim}(z_{x_s}, z_{x}) / \tau)},
\]

where \( \text{sim}(u, v) = u^T v / \|u\| \|v\| \) denotes the pairwise content similarity between feature \( u \) and \( v \) and \( \tau \) 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 \( \tau \) 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
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.

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
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 $1024 \times 320$ 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.

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 for stable training. 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, negative transfer, this does not happen at scale, and the CDCA module could progressively yield more effective source-reference pairs for stable training.

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-
resonances. 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.

**REFERENCES**


