Robust Perception Under Adverse Conditions for Autonomous Driving Based on Data Augmentation

Ziqiang Zheng, Yujie Cheng, Zhichao Xin, Zhibin Yu, *Member, IEEE*, and Bing Zheng, *Member, IEEE*

***Abstract*— Many existing advanced deep learning-based autonomous systems have recently been used for autonomous vehicles. In general, a deep learning-based visual perception system heavily relies on visual perception to recognize and localize dynamic interest objects (*e.g.*, pedestrians and cars) and indicative traffic signs and lights to assist autonomous vehicles in maneuvering safely. However, the performance of existing object recognition algorithms could degrade significantly under some adverse and challenging scenarios including rainy, foggy, and rainy night conditions. The raindrops, light reflection, and low illumination pose a great challenge to robust object recognition. Thus, A robust and accurate autonomous driving system has attracted growing attention from the computer vision community. To achieve robust and accurate visual perception, we target to build effective and efficient augmentation and fusion techniques based on visual perception under various adverse conditions. The unpaired image-to-image (I2I) synthesis is integrated for visual perception *enhancement* and effective synthesis-based *augmentation*. Besides, we design a *two-branch* architecture to utilize the information from both the original image and the enhanced image synthesized by I2I. We comprehensively and hierarchically investigate the performance improvement and limitation of the proposed system based on visual recognition tasks and network backbones. An extensive experimental analysis of various adverse weather conditions is also included. The experimental results have demonstrated the proposed system could promote the ability of autonomous vehicles for robust and accurate perception under adverse weather conditions.**

***Index Terms*— Generative adversarial network, data augmentation, unpaired image-to-image translation.**

I. Introduction

THE development of autonomous driving [1], [2], [3], [4], [5] has greatly benefited from the advancement of computer vision, in which various tasks such as object classification [6], [7], detection [8], [9], [10], semantic segmentation [11] and depth estimation [12], [13] are used for assisting the autonomous driving. The deep learning-based algorithms [14] have achieved promising progress in autonomous driving. Lots of large-scale datasets (*e.g.,* KITTI [15], Apolloscape [16], nuScenes [17], Cityscapes [18], BDD100K [19], Oxford RobotCar [20] datasets, *etc*) have been proposed as benchmarks for measuring the performance of different vision tasks in autonomous driving. The large amounts of data with comprehensive annotations lead to satisfactory visual perception in autonomous driving scenarios.

However, since most of the available annotated datasets were mainly collected in standard conditions, the trained model based on these standard and clear data may suffer from performance degradation under adverse weather conditions (*e.g.,* rainy, foggy, low illumination and rainy night). We refer the readers to check the example images from different autonomous driving conditions in Fig. 1. The visual perception performance of the autonomous driving system is critically degraded under these adverse conditions. A semantic segmentation model trained on the clear images could suffer a 30-40 percent accuracy drop [24], [25] when it is tested on images captured under adverse conditions as illustrated in Fig. 1. Theoretically, sufficient data on various weather conditions can help us to overcome this problem. However, it is time-consuming and expensive to collect such heavy data and provide further annotations, which would result in intensive human labor. To alleviate this, researchers developed physical parameters-based simulators [22], [26], [27], [28] for collecting the data under various weather conditions. However, the simulated images critically suffer from the low naturalness and the domain gap with the real data, which pose a huge challenge for adopting the trained model in the synthetic data into real-world scenarios.

Recent domain adaptation algorithms [29], [30], [31], [32], [33], [34] achieved remarkable success and an overwhelming tendency is to apply domain adaptation to boost recognition performance under adverse conditions. The localization [35], semantic segmentation [36], [37], [38], [39], [40], and object detection [41], [42], [43] have been conducted under the adverse conditions and we could achieve performance gain by integrating the domain adaptation into the whole framework. Chen et al. [41] proposed to perform cross-domain object detection by aligning the feature-level distribution between source and target domains. Though feature-level domain adaptation [32], [33], [37], [38], [41] could help extract

Manuscript received 9 May 2022; revised 31 December 2022; accepted 12 July 2023. Date of publication 18 October 2023; date of current version 29 November 2023. This work was supported in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2021LZH005, in part by the National Natural Science Foundation of China under Grant 62171419, and in part by the Finance Science and Technology Project of 630 Hainan Province of China under Grant ZDKJ202017. The Associate Editor for this article was S. S. Nedevschi. *(Corresponding author: Zhibin Yu.)*

Ziqiang Zheng, Yujie Cheng, and Zhichao Xin are with the School of Electronic Information Engineering, Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266520, China.

Zhibin Yu and Bing Zheng are with the School of Electronic Information Engineering, Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266520, China, and also with the Key Laboratory of Ocean Observation and Information of Hainan Province, Sanya Oceanographic Institution, Ocean University of China, Sanya 572025, China (e-mail: yuzhibin@ouc.edu.cn).

Digital Object Identifier 10.1109/TITS.2023.3297318

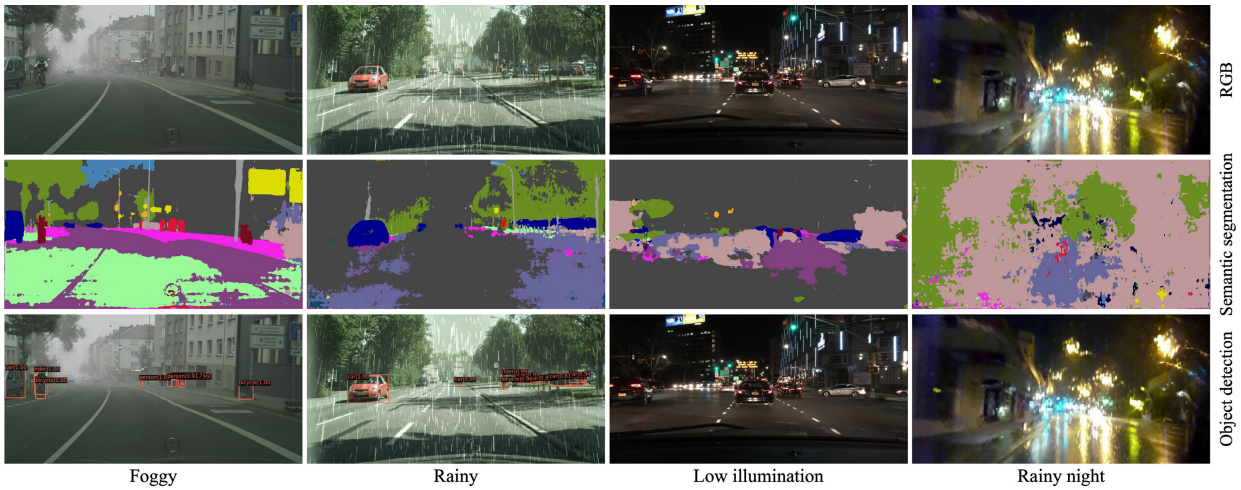

Fig. 1. Example images under adverse weather conditions: foggy image (with 150m visibility) from Foggy Cityscapes [21], rainy image (with 200mm raindrops) from Rainy Cityscapes [22], low illumination image from BDD100K [19] and rainy night image from Alderley [23]. The semantic segmentation and object detection results are also reported. We observe a significant performance drop for both semantic segmentation and object detection under adverse conditions.

the domain-agnostic feature representations from the target images, this line of methods cannot provide the reasonable visibility enhanced outputs for human drivers, which could provide the interaction between humans and autonomous driving system when encountering the corner cases.

Another domain adaptation strategy is to perform image-level domain adaptation between adverse conditions and standard conditions. The existing algorithms have demonstrated superiority in enhancing human visibility [44], [45], [46] based on the paired data. However, it is nearly impossible to collect the paired data under the autonomous driving setting due to the dynamic scenes and objects. The significant progress of unpaired image-to-image (*I2I*) synthesis brings an elegant and effective solution for performing image-level domain adaptation in the autonomous driving setting. The representative CycleGAN [29] provides a generic and elegant solution for unpaired image synthesis based on cycle-consistency loss. The fair generalization ability and less difficulty in data acquisition make Cycle-GAN algorithms [43], [47], [48], [49], [50], [51], [52] popular in the autonomous driving scenarios. However, the great challenges posed by the adverse conditions also lead to a dilemma for the unpaired I2I. The translation model will introduce obvious visual artifacts and unnatural noise after the image synthesis even if the model could enhance the domain-agnostic feature representations meanwhile, causing the *flawed* domain adaptation as described in Fig. 2. In this paper, we target to explore and push the optimal boundaries on adopting the unpaired I2I synthesis as an effective *synthesis-based data augmentation*. We aim to conduct an in-depth and comprehensive analysis of the effectiveness and limitations of adopting the I2I synthesis in autonomous driving scenarios.

In detail, we hierarchically dissect the effectiveness of adopting the unpaired I2I synthesis for generating images in both directions (from a relatively adverse condition to a relatively standard condition and vice versa). The former adverse→standard adaptation (named *Enhancement*) described in Fig. 3 (a) is to consider the synthesized image

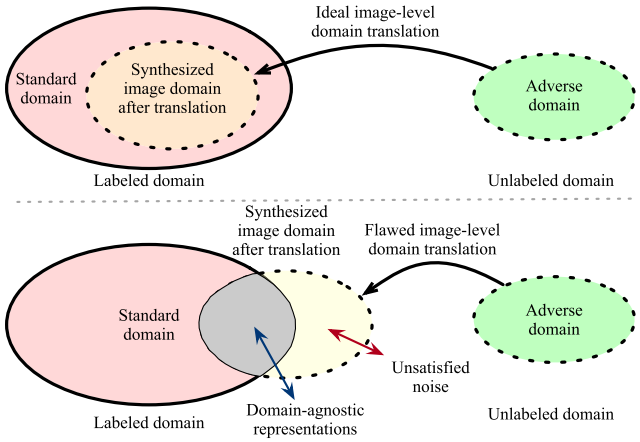

Fig. 2. An ideal I2I synthesis can transfer the adverse domain to the standard domain and not introduce unsatisfactory noise or unnatural objects. The flawed domain translation can enhance the domain-agnostic feature representations and would inevitably introduce visual artifacts, incorrect color, noise, and uncertainty that might decrease the performance of visual perception tasks.

as an enhanced output [35], [43] for promoting the downstream recognition performance. The synthesized images are directly utilized for recognition without retraining the model. The second pipeline (named *Augmentation*) is to perform standard→adverse adaptation for effective data augmentation. The trained translation model is utilized for generating reasonable counterparts while the corresponding labels from the standard domain are preserved to formulate the image-label pairs in the adverse domain. Then we retrain the model (optimized by the data from the standard domain) based on both the standard and generated images for extracting the domain-agnostic feature representations as explained in Fig. 3 (b). However, no matter which route (standard→adverse or adverse→standard) we choose, a significant problem is that some clues from the original images will be discarded and some unsatisfactory noise will be introduced. We argue that both the original images and the synthesized images could

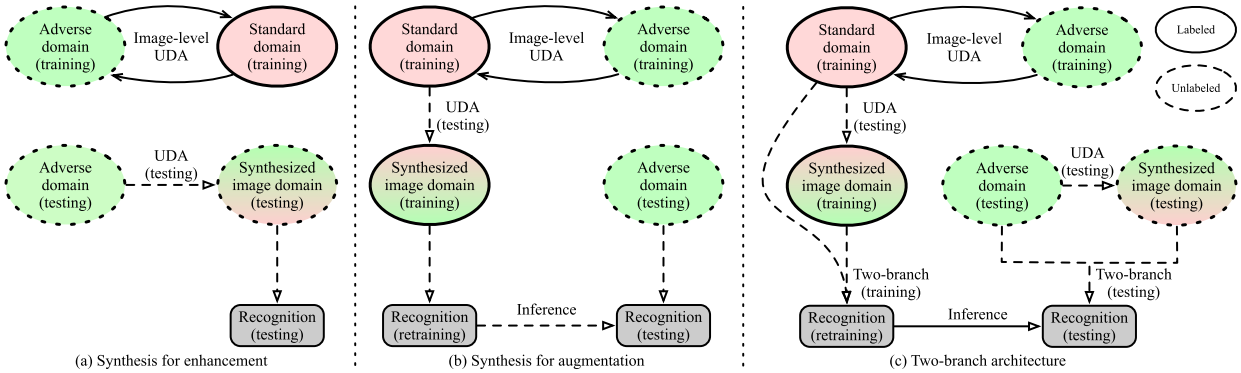

Fig. 3. The illustrations of the proposed three hierarchical settings. a) Synthesis for *enhancement*. We perform unpaired I2I between the adverse domain and standard domain and adopt adverse→standard direction to conduct the synthesis for enhancement. The pre-trained object recognition model based on standard data is only used for evaluation. b) Synthesis for *augmentation*. We perform standard→adverse domain translation for data augmentation. The synthesized images and inherited annotations are utilized to formulate the image-label pairs for retraining/fine-tuning the pre-trained model. Then the real images from the adverse domain are adopted for evaluation based on the fine-tuned model. c) *Two-branch* architecture. Both the real images and the synthesized images after domain translation are utilized to re-train and fine-tune the pre-trained recognition model. Best viewed in color.

contain valuable information. How to alleviate the information loss after the image translation and maximize the efficiency of the synthesis-based data augmentation is critical for promoting further recognition performance. Thus, we design a two-branch training strategy (named *Two-branch*) to take advantage of both the original real images and the synthesized images shown in Fig. 3 (c). The proposed method is also a general training strategy, which could be extended to different translation network backbones and various downstream vision perception tasks. In a nutshell, our main contributions can be summarized as follows:

- We adopt to integrate the unpaired I2I synthesis as an effective synthesis-based data augmentation for promoting visual perception performance under various adverse conditions. We have comprehensively and hierarchically investigated the effectiveness and limitations of synthesis-based data augmentation under various conditions.
- We introduce a general two-branch architecture to make efficient use of the information from both the original images and the enhanced counterparts after the image synthesis. Various downstream recognition tasks and networks and vision recognition tasks are conducted to demonstrate the effectiveness and generalization ability of the proposed two-branch architecture.
- We conduct a direct comparison between *Enhancement*, *Augmentation* and *Two-branch* settings to provide more in-depth analysis. The performance of semantic segmentation, object detection and depth estimation under various adverse conditions can be significantly promoted on various datasets.

The rest part is organized as below. Section II briefly introduces the related work and Section III elaborates the proposed approach. Section IV presents the extensive experimental results on various datasets, followed by the conclusion in Section V.

II. Related Work

In this section, we first review some traditional rule-based data augmentation techniques that are widely adopted in various computer vision tasks. Then some deep learning-based data augmentation algorithms are further investigated. We also include the recent progress of image-to-image synthesis. Furthermore, we discuss synthesis-based data augmentation followed by the scene perception and understanding applications required in autonomous driving.

A. Data Augmentation

Rule-based. Insufficient data is always a huge challenge for data-driven models. The early rule-based data augmentation approaches for images mainly focus on geometric transformations [53], which include random flipping, cropping, rotation, noise injection, *etc*. Similar to geometric transformation, color transformation is another common way to increase the diversity of original data in color space. These data augmentations are carefully designed using the domain knowledge about images. Dropout [54] is one representative regularization method to prevent neural networks from overfitting. The Mixup [55] targets to fuse two images with a random ratio to generate a new image. Cutout [56] randomly covers a region of an input image with a square. Cutmix [57] crops a part region and puts the cropped region in another image to achieve data augmentation. Furthermore, the random erasing operation [58] is also designed to help the network to extract efficient feature representations. These approaches are simple and efficient and many popular deep learning models still consider augmentation transformations [6], [9], [59] as one necessary component to promote recognition performance.

Deep learning-based. Different from the geometric and color transformations, which focus on the input space with low-dimensional representations, the success of deep learning makes possible semantic augmentation [60], [61], [62], [63], [64]. The generative model [64], [65] is an attractive deep learning-based strategy for data augmentation. By exploring the high-dimensional latent space, deep learning methods could perform interpolation and extrapolation to conduct effective data augmentation.

B. I2I Synthesis

The I2I synthesis models can mainly fall into two categories: *paired* [66], [67] and *unpaired* while we mainly focus on the unpaired I2I synthesis models because of the intrinsic superiority of data acquisition. Many GAN variants [29], [35], [48], [51], [68], [69], [70], [71], [72] have been proposed for unpaired I2I synthesis tasks without corresponding image pairs for training. For instance, CycleGAN [29] proposed a generic and elegant solution for unpaired I2I synthesis. The cycle-consistency loss provides a content constraint to regularize the image translation. UNIT [68] added a shared latent space assumption and enforced weight sharing between the two generators to learn a shared common feature space. To improve the diversity of generated results, models such as MUNIT [48], DRIT [49], [50] were proposed to better decompose visual information into domain-agnostic content and domain-specific feature representations. The further StarGAN [73] was developed to achieve the translation between multiple domains by combining an additional classification loss.

C. Synthesis-Based Data Augmentation

With the introduction of generative adversarial networks (GANs) by Ian Goodfellow et al., GANs have shown a superior ability in image synthesis. The further development of the conditional GAN [66], [67], [74] and Cycle-GAN framework [29], [48], [49], [50], [75] expand the applications of GAN-based image synthesis. The high-quality generated images lead to an irresistible trend of synthesis-based data augmentation [43], [47]. The inappropriate augmentation approach may cause an unnecessary bias and even ultimately lead to counter-productive [76]. Unlike the rule-based augmentation approaches, synthesis-based data augmentation is about how to enhance the visual signals without catastrophic information loss and generate more diverse and effective samples [77]. The few-shot image synthesis [78], [79] has demonstrated the ability to generate reasonable images and promote the performance of visual recognition. Through reasonable and effective synthesis-based data augmentation, the robustness of the recognition model to the noise and adversarial attack could be promoted. Besides, the model could also benefit from a strong generalization ability through data augmentation. Shen et al. [63] proposed to interpret and do the semantic face editing in the latent space. Romera et al. [80] proposed to adopt the unpaired image-to-image synthesis to promote visual perception performance under adverse conditions. Reference [80] either regards the adverse-to-normal image translation for image pre-processing or performs the normal-to-adverse for data augmentation, leading to a sub-optimal solution since the image translation inevitably introduces visual artifacts. While most of the I2I synthesis methods only focus on the image qualities and translation efficiency, how to utilize the generated images to boost the performance of vision tasks such as image recognition, semantic segmentation, and object detection is still a challenge [5].

D. Scene Understanding in Autonomous Driving

Object detection, semantic segmentation and depth estimation are fundamental tasks of autonomous driving scene perception and understanding [81]. Adverse weather and undesirable illumination conditions pose challenges to these two vision tasks [2], [81], [82]. For the semantic segmentation under adverse conditions, Porav et al. [31] proposed to adopt lightweight adapters to transform images of different weather and lighting conditions into an ideal condition, in which the transformed images are evaluated by the off-the-shelf models. Representative cross-domain object detection work is from He et al. [59], in which the authors developed a multi-adversarial Faster-RCNN framework for domain-adaptive object detection in driving scenarios. The source and target domain pairs involve regular and foggy Cityscapes, synthetic and real data from two different driving datasets with similar weather conditions. AugGAN [47] aims to combine an image parsing network to enhance object detection performance in nighttime images through day-to-night translation on synthetic datasets. However, it requires paired auxiliary annotations (*e.g.*, semantic segmentation annotations) to regularize the image parsing network. The auxiliary annotations sometimes are expensive to acquire. The mono depth estimation [12], [13] is to obtain the dense depth map from the input sequences in an unsupervised manner. However, the adverse conditions result in a huge challenge for performing reasonable depth estimation or capturing reliable geometric correspondences. Our work addresses how to achieve robust and accurate and robust scene understanding (including object detection, semantic segmentation and depth estimation) under adverse conditions.

TABLE I
LIST OF SYMBOLS AND ABBREVIATION

Notation	Description
$\mathbb{X}$	Real image domain under *adverse* condition
$\mathbb{Y}$	Real image domain under *standard* condition
x, y	Real image sample under *adverse* and *standard* condition
$\tilde{\mathbb{X}}$	Synthesized image domain under *adverse* condition
$\tilde{\mathbb{Y}}$	Synthesized image domain under *standard* condition
$\tilde{x}/\tilde{y}$	Synthesized image sample under *adverse*/*standard* condition
$\mathcal{F}/\mathcal{G}$	Forward and backward functions between $\mathbb{X}$ and $\mathbb{Y}$
Φ_s/Φ_d	Frozen segmentation/detection model trained with Y
$\tilde{\Phi}_s/\tilde{\Phi}_d$	Retrained segmentation/detection model after fine-tuning
$\mathcal{S}$	Shallow CNN module to combine information from two images
Original	Images from $\mathbb{X}$ are tested based on Φ_s/Φ_d
Enhancement	Images translated from $\mathbb{X}$ to $\mathbb{Y}$ are tested based on Φ_s/Φ_d
Augmentation	Translating image from $\mathbb{Y}$ to $\mathbb{X}$ while preserving labels for augmentation
Two-branch	Original and translated images are utilized for training and testing
Oracle	Training and testing in $\mathbb{X}$ with given both labels

III. APPROACH

A. Preliminary

We first illustrate the list of symbols and abbreviations mentioned in our paper in Table I to provide better readability. $\mathbb{X}$ and $\mathbb{Y}$ represent the image domains under the *adverse* (*e.g.*, foggy, rainy, low illumination and rainy night) and the *standard* condition, respectively. The proposed system mainly contains two procedures: *data augmentation* for robust feature extraction and *cross-domain visual perception*. The former stage contains 1) rule-based data transformations and 2) image synthesis for domain adaptation. The rule-based data transformations including random flipping, random cropping, normalization, and color jitter are integrated with the visual perception network for promoting the recognition robustness.

The image synthesis for domain adaptation is to perform the image-level domain adaptation, translating images from one domain to another domain with desired appearance representations through the unpaired I2I. To comprehensively investigate the effectiveness and also the limitations of utilizing the synthetic images for promoting visual perception performance under various conditions, we have designed five settings to provide the hierarchical analysis:

1) *Original* means that we directly perform visual perception on the target images collected under the adverse conditions based on the pre-trained model optimized by image-label pairs from the standard domain.
2) *Enhancement* indicates that we first perform adverse→standard image synthesis and then perform visual perception on the enhanced images based on the pre-trained model from the standard domain.
3) *Augmentation*. We perform standard→adverse domain adaptation and generate the synthetic counterparts in the adverse domain while preserving the label annotations for the source images. Both the source images and the synthetic images are used for further optimizing the visual perception model to make the model domain-agnostic.
4) *Two-branch*. We utilize both original and synthetic images together through parallel branches to perform feature fusion to obtain more robust and accurate feature extraction. Through the two-branch architecture, we can perform visibility enhancement and alleviate the influence caused by unsatisfactory image synthesis meanwhile.
5) *Oracle*. We assume that the image-label pairs are available in the adverse domain and we train the visual perception model from scratch. We finally evaluate visual perception performance based on such a trained model.

Before going into an in-depth explanation, we first go through some basic knowledge about synthesis-based data augmentation. We leave more descriptions about the *Original* and *Oracle* settings in Sec. IV.

B. Image Synthesis for Domain Adaptation

To achieve the pixel-level domain adaptation between two visual image domains: $\mathbb{X}$ and $\mathbb{Y}$ (*e.g.*, the adverse foggy image domain and the standard daytime image domain respectively). Given two images x and y from $\mathbb{X}$ and $\mathbb{Y}$ separately, two reverse functions are introduced: the forward translator $\mathcal{F}$ aims to generate $\tilde{y} = \mathcal{F}(x)$ in the target domain $\mathbb{Y}$ while the reverse translator $\mathcal{G}$ targets to generate the counterpart reconstruction of $\hat{x} = \mathcal{G}(\tilde{y})$ in the original source domain $\mathbb{X}$. To make $\tilde{y}$ as close to y as possible, the widely used adversarial loss is adopted:

$$\mathcal{L}_{adv}(\mathcal{F}, D_{\mathbb{Y}}) = \mathbb{E}_{y\sim\mathbb{Y}}[\log D_{\mathbb{Y}}(y)] + \mathbb{E}_{\tilde{y}\sim\tilde{Y}}[\log(1 - D_{\mathbb{Y}}(\tilde{y}))], \quad \text{with } \tilde{y} = \mathcal{F}(x), \tag{1}$$

where $D_{\mathbb{Y}}$ is the domain-specific discriminator for the domain $\mathbb{Y}$. The reverse adversarial loss $\mathcal{L}_{adv}(\mathcal{G}, D_{\mathbb{X}})$ is also computed to help generate reasonable image outputs, where D_x is the domain-specific discriminator for domain $\mathbb{X}$. At the training stage, x and y are randomly selected from $\mathbb{X}$ and $\mathbb{Y}$. In other words, x and y are not paired. To preserve the content information, the cycle-consistency loss $\mathcal{L}_{cyc}$ [29] is adopted to link F and G:

$$\mathcal{L}_{cyc}(\mathcal{F}, \mathcal{G}) = \mathbb{E}_{x\sim\mathbb{X}}[\|\mathcal{G}(\mathcal{F}(x)) - x\|_1] + \mathbb{E}_{y\sim\mathbb{Y}}[\|\mathcal{F}(\mathcal{G}(y)) - y\|_1], \tag{2}$$

through the pixel-wise distance, we can preserve the content information after the unpaired image-to-image translation. Please note that the synthesized images are task-agnostic, which indicates that the synthetic images can support various downstream visual perception tasks (*e.g.*, object detection, semantic segmentation and depth estimation in this work).

C. Synthesis for Enhancement

Φ_s and Φ_d denote the downstream frozen segmentation and detection models trained on the standard clear images. Ideally, we can enhance the recognition-relative signals while only a few noises or visual artifacts introduced. Thus, we could directly perform the visual recognition and compute the prediction accuracy based on the synthetic output $\tilde{y}$. To be noted, the two procedures (the synthesis-based data augmentation models ($\mathcal{F}$ and $\mathcal{G}$) and the downstream vision recognition networks Φ_s and Φ_d) are optimized separately with the following loss functions:

$$\mathcal{L}_{seg}(\Phi_s, y) = -\mathbb{E}_{y\sim\mathbb{Y}}[\frac{1}{N_{seg}}\sum_{i=1}^{N_{seg}}\sum_{c=1}^{C} g_i^c \log s_i^c] \tag{3}$$

$$\mathcal{L}_{det}(\Phi_d, y) = -\mathbb{E}_{y\sim\mathbb{Y}}[\frac{1}{N_{det}}\sum_j \mathcal{L}_{cls}(p_j, p_j^*) + \lambda\frac{1}{N_{reg}}\sum_i p_j^*\mathcal{L}_{reg}(t_j, t_j^*)] \tag{4}$$

where $\mathcal{L}_{seg}$ denotes the pixel-wise cross entropy classification loss for semantic segmentation and $\mathcal{L}_{det}$ is the total loss (including the classification loss $\mathcal{L}_{cls}$ and the regression loss $\mathcal{L}_{reg}$) for the detection task. λ is the hyper-parameter to balance the contribution of the two parts. N_{reg} and N_{det} are adopted for achieving the normalization. g_i^c is a binary indicator if class label c (C is the number of total class) is the correct classification for pixel i, and s_i^c is the corresponding predicted probability for segmentation. j denotes the index of an anchor in a mini-batch; p_j and p_j^* are the predicted probability and the corresponding ground truth label of anchor j being an object. Similarly, t_j and t_j^* are the predicted parameterized coordinates and the corresponding ground truth of the predicted bounding box for anchor j.

When obtaining the optimized Φ_s and Φ_d, we adopt $\mathcal{F}$ to translate the images captured in the adverse domain to the standard domain. We observe that there is a significant performance improvement on $\Phi_s(\mathcal{F}(x_{test}))$ and $\Phi_d(\mathcal{F}(x_{test}))$ by translating the adverse foggy images into the standard clear images. However, this would not work for all the conditions. For the extremely low illumination images in BDD100K [19] dataset, Φ_d would under-perform to recognize the translated

images. We attribute this performance drop to the noise and uncertainty caused by the unpaired image-to-image translation model.

D. Synthesis for Augmentation

The synthesis for enhancement and the synthetic outputs evaluated on the frozen models will result in a performance drop under the low illumination condition. To address this issue, we try to adopt the synthesis to achieve data augmentation. In detail, we perform the $\mathbb{Y} \rightarrow \mathbb{X}$ translation since it is much easier than the $\mathbb{X} \rightarrow \mathbb{Y}$ discussed in [52]. After the translation, we preserve the labels of $\mathbb{Y}$ for training since it is easier to collect data under the standard condition and do the corresponding annotations. Then we finetune the Φ_s/Φ_d to obtain $\tilde{\Phi}_s/\tilde{\Phi}_d$ using the synthetic outputs and preserved labels to force the model to learn the domain-specific feature representations.

$$\mathcal{L}_{seg}(\tilde{\Phi}_s, \tilde{x}) = -\mathbb{E}_{\tilde{x}\sim\tilde{\mathbb{X}}}[\frac{1}{N_{seg}}\sum_{i=1}^{N_{seg}}\sum_{c=1}^{C} g_i^c \log s_i^c], \tag{5}$$

$$\mathcal{L}_{det}(\tilde{\Phi}_d, \tilde{x}) = -\mathbb{E}_{\tilde{x}\sim\tilde{\mathbb{X}}}[\frac{1}{N_{det}}\sum_{j}\mathcal{L}_{cls}(p_j, p_j^*) - \frac{\lambda}{N_{reg}}\sum_{i} p_j^*\mathcal{L}_{reg}(t_j, t_j^*)], \tag{6}$$

please note, at the testing stage, the real nighttime images are evaluated rather than the synthesized images. The retrained models $\tilde{\Phi}_s$ and $\tilde{\Phi}_d$ can accurately recognize the objects under some challenging adverse conditions: $\tilde{\Phi}_s(x_{test})$ and $\tilde{\Phi}_d(x_{test})$. However, there is still a drawback that the original real images are discarded during the retraining procedure. Under this setting, we cannot fully utilize the information from real images and the correspondences between real images and synthesized counterparts.

E. Two-Branch

To further better utilize the information from both original and synthesized image outputs, we design a two-branch architecture, which can fuse the information from different inputs to promote the recognition ability of models. Given the original image x and the translated enhanced image output $\tilde{y}$, we design a shallow CNN network $\mathcal{S}$ (including E_r and E_s responsible for extracting information from the original real images and translated images respectively) shown in Fig. 4 to fuse information. Since both E_r and E_t contain very shallow CNN modules (only contains one residual block to achieve information residual extraction), we would **not** introduce lots of computational costs and network parameters. We then concatenate the feature representations $E_r(x)$ and $E_s(\tilde{y})$ to do the feature fusion. We inherit the network backbone from the vanilla recognition backbone for the following network. The loss function can be rewritten as follows:

$$\mathcal{L}_{seg}(\tilde{\Phi}_s, \mathcal{S}(x, \tilde{y})) = -\mathbb{E}_{x\sim\mathbb{X},\tilde{y}\sim\tilde{\mathbb{Y}}}[\frac{1}{N_{seg}}\sum_{i=1}^{N_{seg}}\sum_{c=1}^{C} g_i^c \log s_i^c] \tag{7}$$

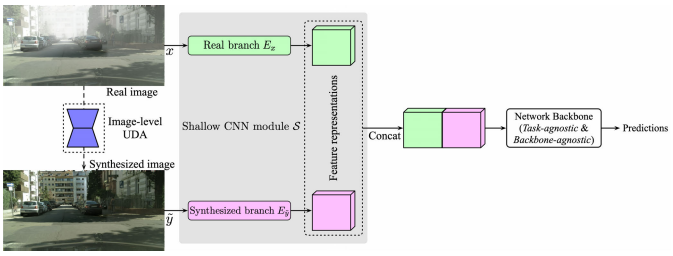

Fig. 4. The designed two-branch architecture. We perform the image-level unsupervised domain adaptation to obtain the corresponding synthesized outputs. E_r and E_t contain very shallow CNN modules $\mathcal{S}$ for feature extraction.

$$\mathcal{L}_{det}(\tilde{\Phi}_d, \mathcal{S}(x, \tilde{y})) = -\mathbb{E}_{x\sim\mathbb{X},\tilde{y}\sim\tilde{\mathbb{Y}}}[\frac{1}{N_{det}}\sum_{j}\mathcal{L}_{cls}(p_j, p_j^*) - \frac{\lambda}{N_{reg}}\sum_{i} p_j^*\mathcal{L}_{reg}(t_j, t_j^*)]. \tag{8}$$

Then, we retrain the whole model based on the two-branch architecture. To be noted, the proposed two-branch architecture is a general framework, which could be extended to various recognition backbones. Through the two-branch architecture, we can build a complementary system to utilize the lossless information from the original real image and the enhanced visual signals from the translated image to promote recognition performance.

IV. Experimental Results

A. Experimental Setup

1) Datasets: **Cityscapes** [18] is a large-scale dataset consisting of diverse urban street scene videos captured across 50 different cities at varying times of the year. This dataset provides the corresponding semantic segmentation and object detection annotations for each image.

Foggy Cityscapes [21] is a recently proposed synthetic foggy dataset simulating fog on real scenes with three different levels of visibility: 150m, 300m, and 600m visual visibility. For this dataset, we choose clear images and foggy images with 600m visibility following the original split. In all the experiments, 2975 clear and corresponding foggy images are adopted for training, and the other 500 clear and corresponding foggy images for evaluation.

Rainy Cityscapes [22] is rendered from the physics-based model and it provides 7 rain levels (from dizzle to storm conditions). We adopt the most challenging images with 200mm raindrops in our experiments. The image resolution is 1024×512. Similarly, 2975 clear and rainy images are adopted for training, and the other 500 clear and rainy images for evaluation.

Alderley is originally proposed for the SeqSLAM algorithm [23], which collected the images for the same route twice: one on a sunny day and one on a stormy rainy night. Every frame in the dataset is GPS-tagged, and thus each nighttime frame has a corresponding daytime frame. We use the GPS-matching annotation for evaluation.

SeeingThroughFog [83] is collected under various conditions including rainy, snowy and foggy images. This dataset

provides the 2D bounding box annotation for about 12,000 samples. The information streams from different multi-modal sensors (*e.g.*, stereo camera, gated camera, Radar, Lidar, FIR camera and IMU) are collected for feature fusion to promote visual perception performance under adverse conditions.

BDD100K [19] is a large-scale high-resolution autonomous driving dataset, which contains 100,000 video clips collected from various cities under different conditions. It provides various annotations for the keyframe from each video clip, including bounding box annotations. We reorganized this dataset according to the annotation of the daytime data and obtained a 27,971/3,929 train/test split for night images and a 36,728/5,258 train/test split for day images.

2) Evaluation Metrics: **Semantic segmentation**. Intersection-Over-Union (IoU) is a commonly used metric for semantic segmentation. For each object class, the IoU is the overlap between predicted segmentation and given ground truth, divided by the union of the prediction and ground truth. In the case of multiple classes, we take the average IoU of all classes (mIoU) and the per-class IoU results. For semantic segmentation, we adopt the DeepLabv3+ [84] model on Cityscapes dataset [18] for evaluation.

Object detection. We use mean average precision (mAP): AP_{95} and AP_{50} to evaluate the performance of object detection. We also report the average precision for individual classes to have a more thorough view of the detection performance under various experimental settings.

Depth estimation. We compute the error and accuracy between the predicted depth estimation outputs and given ground truths following the setting of [13]. The RMSE, RMSE(log), Abs Rel and Sq Rel are computed (lower is better). The accuracy under three settings: $\delta < 1.25, \delta < 1.25^2, \delta < 1.25^3$ are computed (higher is better).

3) Implementation Details: We perform the unpaired I2I translation at the image resolution of 1024×512. The generator architecture is borrowed from CycleGAN [29][1] with some modifications. Four Conv-InstanceNorm-ReLU layers are used to generate the intermediate feature representations, the kernel size is set to 3 and the stride is 2. For the bottleneck, 9 residual blocks are applied to stack the content information. 4 inverse Deconv-InstanceNorm-ReLU blocks are performed to generate the same size image outputs. The Tanh activity function is applied to obtain the normalized outputs. To guarantee the naturalness and the image quality of generated images and reduce the domain gap, we adopt the multi-scale discriminator architecture [48]. The hinge-based adversarial loss is applied for more stable adversarial training. The feature matching loss designed in Pix2pixHD [67] is also included to match the intermediate feature representations. After the training, the synthetic images are resized to the original size through bilinear interpolation for downstream visual perception. The batch size is set to 16 for optimizing the object detector (8 for the semantic segmentation model). Both the original images and the synthesized counterparts with label annotations are used for training visual perception models. For optimizing the unpaired image synthesis model, the batch size is set to 2. To improve the generality and robustness of models, we use some common rule-based data augmentation approaches including random flipping, random resize and center crop. We adopt Adam [85] to optimize our model.

Fig. 5. The weighted semantic segmentation results of the original images and the enhanced images after image synthesis under foggy and rainy conditions. The semantic segmentation outputs are obtained by the pre-trained DeepLabv3+ [84] model.

We also provide the implementation details of utilizing synthesized images to promote visual perception performance under various adverse conditions. The FPS (frame-per-second) of the synthesizing desired image is **4.63** with image resolution 1024×512 when tested on the single GTX Geforce 3090. After synthesis, the generated images are resized to the original image size (*e.g.*, 2048 × 1024 for Foggy Cityscapes dataset and 1280 × 720 for BDD100K dataset) through the bilinear interpolation. We conduct the downstream object detection and semantic segmentation tasks based on MMDetection[2] and MMSegmentation[3] respectively. Under the *Enhancement* setting, the synthesized images after the adverse-to-normal adaptation are directly evaluated based on the pre-trained model on the normal condition. For *Augmentation*, both the original images and the synthesized counterparts with label annotations are used for training visual perception models. We modified the data loader codes of MMDetection and MMSegmentation to guarantee that the two parallel branches are using the same data augmentation to ensure consistent feature representations under the *Two-branch* setting.

B. Simulated Rainy and Foggy

We first conduct experiments on two physically simulated datasets: Foggy Cityscapes dataset [21] and Rainy Cityscapes dataset [22], which are both simulated from the Cityscapes dataset [18].

1) Synthesis for Enhancement: We first explore the performance improvement introduced by synthesis-based enhancement. Given the pre-trained model: DeepLabv3+ [84] model for semantic segmentation and Faster-RCNN [8] model for object detection optimized by the standard clear images, the images collected under the adverse condition (*e.g.*, foggy and rainy) are evaluated based on the frozen model. We first conduct the semantic segmentation on the Foggy Cityscapes dataset [21] (150m visibility) and Rainy Cityscapes (200mm

[1]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[2]https://github.com/open-mmlab/mmdetection

[3]https://github.com/open-mmlab/mmsegmentation

TABLE II

THE SEMANTIC SEGMENTATION PERFORMANCE COMPARISON ON FOGGY CITYSCAPES DATASET [21] WITH 150M VISUAL VISIBILITY AND RAINY CITYSCAPES DATASET [22] WITH 200MM RAINDROPS. THE BEST RESULTS ARE IN BOLD

Exp	Method	Setting	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU↑
1	Original	Foggy Cityscapes	97.6	81.7	77.8	46.8	53.3	58.7	59.1	72.7	71.5	55.5	60.2	77.8	**61.0**	91.3	71.0	76.9	65.1	64.9	74.6	69.3
2	CycleGAN		96.3	75.3	77.5	29.4	33.6	42.8	49.2	59.7	77.6	44.1	69.6	72.5	51.7	90.6	69.4	72.8	59.0	47.1	68.2	62.4
3	Enhancement		97.8	82.3	86.8	49.0	54.7	58.3	62.0	71.2	86.6	57.2	86.3	76.6	58.8	92.8	71.9	79.6	64.5	60.4	73.0	72.1
4	Augmentation		97.5	84.1	87.3	54.2	58.3	59.5	64.2	73.1	87.5	59.4	85.7	77.4	60.2	93.3	74.3	81.8	68.3	64.6	74.1	73.9
5	Two-branch		**97.8**	**84.7**	**88.0**	**56.4**	**59.6**	**60.3**	**65.4**	**74.3**	88.0	**60.0**	85.9	**77.8**	60.6	**94.0**	**74.6**	**82.4**	**69.1**	**65.1**	**77.0**	**74.8**
6	Oracle		96.8	80.0	87.3	35.6	46.8	55.9	56.0	68.0	**89.2**	47.1	**89.4**	73.3	48.2	91.7	42.1	40.5	18.9	38.9	67.7	61.7
7	Original	Rainy Cityscapes	24.3	4.6	37.0	3.3	4.9	9.7	17.4	31.6	22.6	14.9	0.1	30.4	5.3	10.2	5.3	1.0	1.1	0.4	22.8	13.0
8	CycleGAN		26.5	5.7	38.4	7.2	10.2	14.6	16.9	33.2	25.7	16.2	3.2	35.2	7.5	13.6	7.2	4.6	2.6	3.2	28.4	15.8
9	Enhancement		96.0	71.5	74.7	27.7	33.5	38.0	30.2	47.3	80.5	48.9	22.8	57.0	23.9	82.0	29.4	54.8	42.0	18.6	52.3	49.0
10	Augmentation		96.5	74.4	77.4	35.2	38.4	43.2	36.3	52.4	83.2	53.7	26.3	59.5	28.7	84.2	35.2	56.8	44.5	20.4	55.1	52.7
11	Two-branch		**96.9**	75.6	78.5	37.9	39.4	45.7	**39.7**	55.3	84.5	**54.1**	28.5	60.4	30.0	85.2	36.1	57.6	**45.1**	22.0	56.2	54.1
12	Oracle		96.4	**79.8**	**85.7**	**39.2**	**50.2**	**56.3**	6.7	**67.3**	**87.6**	48.6	**58.9**	**74.9**	**50.9**	**91.0**	**36.5**	**57.8**	35.6	**46.4**	**79.0**	**60.0**

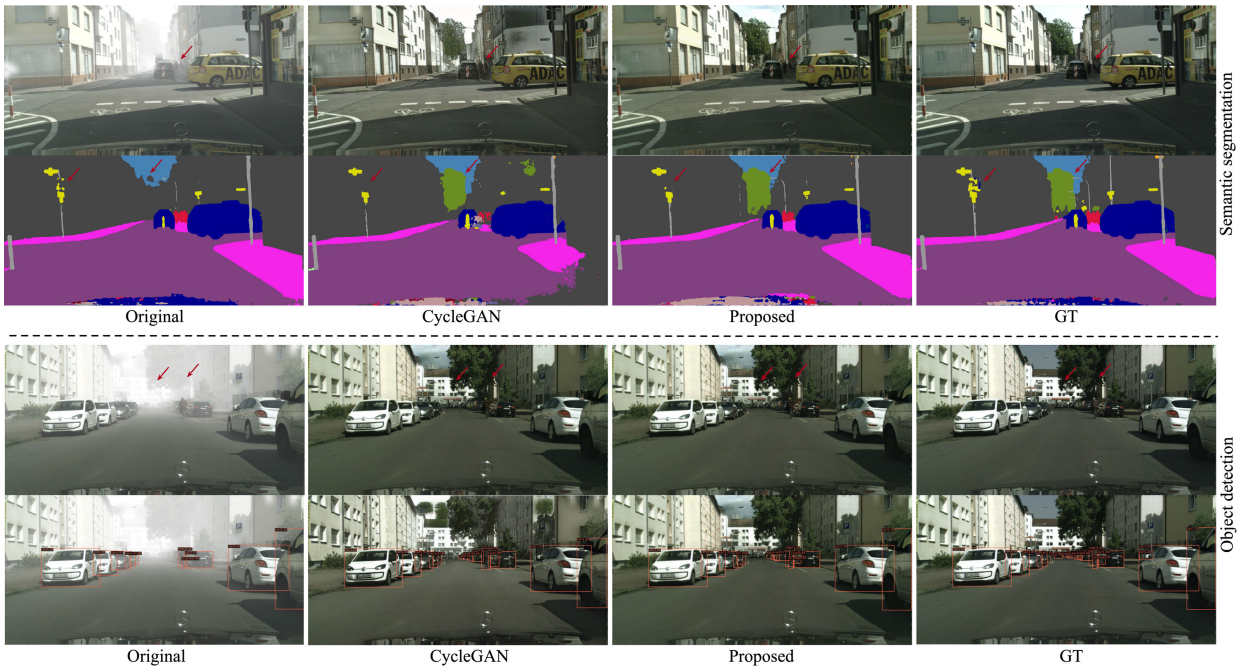

Fig. 6. The visibility, semantic segmentation and object detection result comparison between original foggy images, the enhanced images (generated by CycleGAN and our method) and the clear images. We also provide the ground truths of semantic segmentation and object detection for better comparison. The foggy images with 150m visual visibility are chosen for evaluation. Best viewed in color.

raindrops) dataset. The dense fogginess and raindrops could lead to a huge difficulty to achieve accurate recognition extremely for tiny objects such as traffic signs. The qualitative comparison between the *Original* and *Enhancement* settings is illustrated in Fig. 5. As reported, with the synthesis-based image enhancement, we could obtain more accurate semantic segmentation results for both foggy and rainy conditions. The quantitative results under the two settings (Exp 1 & 3 and Exp 7 & 9) are also reported in Table II. Following the same way, we also perform the cross-domain object detection and report the quantitative results of both *Original* and *Enhancement* settings (Exp 1 & 3 and Exp 7 & 9) in Table III. Besides, to make an intuitive comparison, we also adopt the representative CycleGAN [29] for performing visibility enhancement. We provide the quantitative semantic segmentation and object detection results of utilizing the CycleGAN for enhancement in Table II (Exp 2 and 8) and Table III (Exp 2 and 8), respectively. As reported, our method could achieve more performance gain than CycleGAN, which indicates a stronger ability for promoting cross-domain visual perception performance. Furthermore, we provide the visibility and also image synthesis quality comparison between the synthesized images generated by CycleGAN and our method in Fig. 6 (we only choose the foggy images for visualization). Meanwhile, the semantic segmentation and object detection results of the corresponding images are also reported. As reported, our method can heavily promote both semantic segmentation and object detection performance by translating the foggy images to clear images without losing the detailed content information. In the contrast, CycleGAN cannot obtain the performance gain or even leads to performance degradation due to the failure to synthesize precise and reasonable image outputs. Especially, our method could achieve much more improvements than CycleGAN in performing cross-domain semantic segmentation on the Rainy Cityscapes dataset by comparing Exp 2 and 3 in Table II.

2) Synthesis for Augmentation: We then perform experiments under the *Augmentation* setting (performing standard→adverse image-level domain adaptation for augmentation) and report the quantitative results in Table II

TABLE III

THE OBJECT DETECTION PERFORMANCE COMPARISON ON FOGGY CITYSCAPES DATASET [21] WITH 150M VISUAL VISIBILITY AND RAINY CITYSCAPES DATASET [22] WITH 200MM RAINDROPS. WE ADOPT THE FASTER-RCNN NETWORK AS THE BACKBONE FOR OBJECT DETECTION

Exp	Method	Setting	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP_{95} ↑	mAP_{50} ↑
1	Original	Foggy Cityscapes	23.0	28.6	28.8	10.0	22.3	4.0	6.8	19.3	17.8	26.9
2	CycleGAN		23.0	28.3	36.9	13.5	28.2	5.8	13.0	20.0	20.8	35.2
3	Enhancement		24.4	27.2	39.4	15.6	34.7	11.1	13.0	20.0	23.2	38.1
4	Augmentation		27.0	31.7	44.7	17.4	34.6	13.0	17.1	24.9	26.3	50.2
5	Two-branch		27.3	32.1	45.0	18.7	**35.4**	**13.5**	16.8	25.3	26.8	50.2
6	Oracle		**29.6**	**32.7**	**47.5**	**21.1**	33.1	10.5	**20.2**	**27.1**	**27.7**	**52.3**
7	Original	Rainy Cityscapes	11.2	8.4	20.6	3.8	5.1	0.6	1.2	7.1	7.2	14.8
8	CycleGAN		14.8	13.5	30.5	5.0	13.6	0.6	2.8	8.1	11.1	23.8
9	Enhancement		17.8	15.3	36.1	13.3	24.0	4.2	7.3	12.8	16.3	32.6
10	Augmentation		19.9	19.2	39.3	14.0	24.6	6.0	8.5	15.4	18.4	36.4
11	Two-branch		20.2	19.9	**39.5**	**14.2**	**25.5**	**6.4**	8.6	**15.5**	**18.7**	**37.3**
12	Oracle		**20.4**	**21.9**	38.2	7.0	21.0	4.0	**10.1**	14.5	17.1	34.6

(Exp 4 and 10) and Table III (Exp 4 and 10) for semantic segmentation and object detection, respectively. Through combining the standard→adverse domain adaptation for augmentation, we could achieve further improvement compared with the *Enhancement* setting for both semantic segmentation and object detection: from 72.1 mIoU to 73.9 mIoU for semantic segmentation on Foggy Cityscapes dataset (from 49.0 mIoU to 52.7 mIoU on Rainy Cityscapes dataset); from 23.2 mAP_{95} to 26.3 mAP_{95} for object detection on Foggy Cityscapes dataset (from 16.3 mAP_{95} to 18.4 mAP_{95} on Rainy Cityscapes dataset). We attribute these performance gains to the reason that the model could better extract the domain-agnostic feature representations by combining both the source standard images and the synthesized target images in the adverse conditions for training.

3) Two-Branch: The paradigm of adopting the unpaired I2I image synthesis for data augmentation has discarded some information from the original real images. Also, the uncertainty and noise could also be introduced by the unpaired I2I synthesis. To address these two issues, we design a *Two-branch* architecture to fuse the information from both the original lossless images and the synthesized enhanced outputs. Similarly, we first perform the standard→adverse image-level domain adaptation to generate the "*flawed*" outputs in the target domain and perform feature fusion from the two parallel inputs. The experimental results are reported in Table II (Exp 5 and 11) and Table III (Exp 5 and 11) for semantic segmentation and object detection, respectively. We can achieve further improvement through the two-branch architecture while introducing ignorable computation costs. During the inference stage, we perform the adverse→standard adaptation as the *Enhancement* setting to obtain the input for the synthesized branch as shown in Fig. 4.

4) Direct Comparison: We finally provide a direct comparison between the above-mentioned four settings plus with the *Oracle* setting, in which we train the semantic segmentation and object detection models based on the image-label pairs in the adverse domain from scratch. All the experimental results are reported in Table II and Table III. Please note that we use the unseen real target images from the adverse domain for evaluation under all the settings. The proposed method could even achieve better results than the *Oracle* setting for cross-domain semantic segmentation task on the Foggy Cityscapes dataset. With both the clear images with perfect visibility and synthesized images with desired target appearance representations, the model is optimized without bias and could better disentangle the domain-agnostic and domain-specific feature representations, thus leading to the performance gain.

5) Backbone-Agnostic: We have also demonstrated that the proposed synthesis-based data augmentation for enhancement is backbone-agnostic. We take the cross-domain object detection on the Foggy Cityscapes dataset as an example. We choose various object detection backbones: Reppoints [10] (anchor-free), Cascade RCNN [86] (two-stage) and Deformable DETR [87] (transformer based) for conducting the corresponding experiments. We report the quantitative results in Table IV. The appropriate image enhancement could achieve the performance gain without retraining the model under foggy conditions. However, this image synthesis for enhancement pipeline cannot work well under some very challenging conditions.

C. Nighttime

1) Direct Comparison: Following the experimental setup in Sec. IV-B.4, we perform the cross-domain object detection experiments on the nighttime images from the BDD100K dataset [19] and all the experimental results are reported in Table V. The *Oracle* setting indicates that we use the 27,971 real nighttime images for training and the unseen 3,929 nighttime images for evaluation. As reported, directly adopting the translated images (from the nighttime domain to the daytime domain) for evaluation based on the pre-trained model on the daytime images can even lead to a significant performance drop (from 22.1 mAP_{95} to 15.8 mAP_{95}). The object detection performance of nearly every object category has dropped, which indicates that the *Enhancement* setting cannot provide a strong guarantee to achieve performance gain under the challenging nighttime condition. The synthesized visual artifacts leave it even worse to detect meaningful objects. Instead, by adopting the day-to-night (also the standard-to-adverse) adaptation for data augmentation, we can promote

TABLE IV
THE OBJECT DETECTION PERFORMANCE COMPARISON ON FOGGY CITYSCAPES DATASET [21] WITH 150M VISUAL VISIBILITY. WE ADOPT VARIOUS OBJECT DETECTION BACKBONES TO CONDUCT EXPERIMENTS

Backbone	Enhancement	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP_{95} ↑	mAP_{50} ↑
Reppoints	×	**24.5**	**27.5**	33.4	**18.6**	25.1	2.7	14.4	**23.1**	21.2	35.1
	✓	23.6	27.5	**38.4**	18.3	**35.2**	**7.7**	**15.2**	21.7	**23.4**	**42.1**
Cascade RCNN	×	22.9	25.4	37.8	**20.1**	29.2	5.2	15.3	**24.1**	23.0	40.3
	✓	**25.0**	**27.3**	**39.7**	18.8	**34.4**	**6.1**	**16.4**	22.0	**23.7**	**41.4**
Deform. DETR	×	21.8	24.7	29.4	8.9	14.5	1.4	10.3	17.2	16.0	26.5
	✓	**25.7**	**27.8**	**41.3**	**16.5**	**35.2**	**11.8**	**18.1**	**20.9**	**24.7**	**43.2**

the ability of the model to recognize the objects under the low illumination condition by comparing Exp 1, 2 and 3. Though the proposed method can boost overall object detection performance under the *Augmentation* setting, we still observe some failure cases and limitations. In detail, our method can heavily improve the recognition performance of static objects (*e.g.,* "car", "bus", "truck", "traffic light" and "traffic sign"). However, the visual perception performance of moving objects such as "person", "rider", "bike", and "motor" has been degraded. Furthermore, we also perform experiments based on the *Two-branch* architecture.

2) Effectiveness of Two-Branch Architecture: We further assume that the detection annotations of the adverse nighttime domain $\mathbb{X}$ are also available. We perform the day-to-night (night-to-day) domain adaptation for daytime (nighttime) images, respectively to double the whole training datasets. We utilize both the real and synthesized images for training the domain-agnostic object detector. Similarly, we adopt various detection backbones to prove our two-branch architecture is backbone-agnostic and effective. We adopt YOLO-V3 [9] (one-stage), Faster-RCNN [8] (two-stage), Reppoints [10] (anchor-free) and Deformable DETR [88] (Transformer-based) to conduct experiments. The detection results are reported in Table VI. Under this setting, we can promote the object detection performance based on the proposed two-branch architecture. Our method can promote detection performance by adopting various detection backbones, which indicates that the proposed method is a general strategy.

D. Further Results

1) Comparison With Domain Adaptive Object Detection Algorithms: We also provide a comparison with the current domain adaptive object detection algorithms. Following the experimental setting of [92], we perform experiments on the Cityscapes → Foggy Cityscapes adaptation task. The average of AP_{50} rather than the AP_{95} is calculated. To provide a fair comparison, the shorter side of the input images (including both the original images and the corresponding synthesized images) is set to 600-pixel length. The qualitative and quantitative results are illustrated in Fig. 7 and Table VII respectively. For comparison, we included the recent competitive DA-faster [41], SCL [89], GPA [90], UMT [91] and SIGMA [92], which are all specially designed for the cross-domain object detection. As reported, the proposed method could achieve the best results among all the algorithms on the Cityscapes→Foggy Cityscapes adaptation. Please note that these algorithms do not focus on pixel-level domain adaptation or generating counterparts in the target domain. They instead perform the feature-level domain adaptation to align the feature distribution between source and target domains.

2) Real-World Foggy, Snowy and Rainy Data: We perform the object detection experiments on the SeeingThroughFog dataset [83]. Follow the same experimental setup of [83], we report the quantitative comparison with Image-only SSD [83], CYCADA [34], ADDA [33] and DEEP FUSION [83] in Table VIII. The Image-only SSD indicates performing the object detection based on single modality image inputs. We perform the clear→foggy/snowy/rainy image-level domain adaptation to perform the data augmentation. Compared with the feature-level domain adaptation algorithms: ADDA and CYCADA, the proposed method could achieve better detection results while our method could not beat DEEP FUSION [83] since it combined the supervision from other modalities for training, which could provide more reliable and accurate input signals than the visual images.

3) Rainy Night: We also perform experiments under various settings on the challenging Alderley dataset [23]. Due to that there are no manually annotated dense pixel-level semantic segmentation annotations, we only provide the qualitative results in Fig. 8 for the Alderley dataset. As reported, without the night-to-day image translation, the pre-trained semantic segmentation DeepLabv3+ model [84] on the Cityscapes dataset cannot obtain meaningful and reasonable segmentation outputs. The segmentation performance could be promoted based on the enhanced images. Besides, we can further promote the segmentation performance by fine-tuning the segmentation model in the adverse domain. By fusing the signals from both daytime and translated nighttime images, we can obtain more reliable and robust visual perception results. Finally, we should also note that the segmentation performance is also subject to the backbone of the segmentation model. We can further promote the segmentation performance by replacing it with a more powerful backbone.

E. Mono Depth Estimation

In this section, we explore the effectiveness of *Enhancement* for the mono depth estimation [13]. We first adopt

TABLE V

THE OBJECT DETECTION PERFORMANCE COMPARISON ON BDD100K [19] DATASET. AT THE TESTING STAGE, WE CHOOSE THE REAL NIGHTTIME IMAGES FOR EVALUATION FOR ALL THE SETTINGS

Exp	Method	Backbone	person	rider	car	bus	truck	bike	motor	traffic light	traffic sign	train	mAP_{95} ↑	mAP_{50} ↑
1	Original	Faster-RCNN	26.2	14.4	37.5	29.9	30.8	18.6	16.4	14.7	33.2	0.0	22.1	44.4
2	Enhancement		17.7	10.1	30.2	25.8	25.4	12.5	9.0	5.6	22.0	0.0	15.8	31.7
3	Augmentation		24.6	12.1	40.4	33.1	32.4	16.2	14.9	17.0	34.1	0.0	22.5	45.9
4	Two-branch		**27.8**	**15.4**	38.4	30.8	31.5	**19.4**	**17.3**	15.3	34.2	0.0	22.8	45.8
5	Oracle		26.7	13.1	**42.1**	**33.9**	**35.1**	16.8	17.0	**18.4**	**36.1**	0.0	**23.9**	**47.5**

TABLE VI

THE OBJECT DETECTION PERFORMANCE COMPARISON ON BDD100K DATASET [19]. WE ADOPT VARIOUS OBJECT DETECTION BACKBONES FOR OBJECT DETECTION. THE IMAGE-LABEL PAIRS FROM BOTH ADVERSE ($\mathbb{X}$) AND STANDARD ($\mathbb{Y}$) DOMAINS ARE AVAILABLE

Backbone	Two-branch	person	rider	car	bus	truck	bike	motor	traffic light	traffic sign	train	mAP_{95}	mAP_{50}
Faster-RCNN	×	32.1	22.9	46.8	42.3	40.7	20.6	21.5	24.6	35.1	0.0	28.1	52.9
	✓	**33.1**	**24.4**	**49.1**	**43.1**	**42.5**	**21.7**	**22.5**	**25.6**	**37.1**	0.0	**28.8**	**55.1**
YOLO-V3	×	**35.6**	**22.1**	**45.8**	44.8	**26.4**	21.1	14.7	25.2	**39.0**	0.0	27.8	48.8
	✓	35.1	21.3	44.8	**45.8**	25.6	**21.6**	**17.3**	**26.1**	38.1	**7.1**	**28.3**	**50.6**
Reppoints	×	33.7	25.4	**51.5**	45.1	42.9	23.1	22.2	**27.1**	36.7	0.0	30.9	56.1
	✓	**35.0**	**25.5**	48.2	**46.8**	**44.7**	**25.6**	**23.1**	26.9	**39.3**	**0.2**	**31.5**	**58.3**
Deform. DETR	×	33.2	22.0	45.6	44.1	42.2	21.9	21.4	25.1	37.6	0.0	29.3	55.4
	✓	**33.6**	**22.4**	**46.4**	**44.4**	**42.5**	**22.6**	**21.9**	**25.5**	**38.3**	0.0	**29.8**	**56.2**

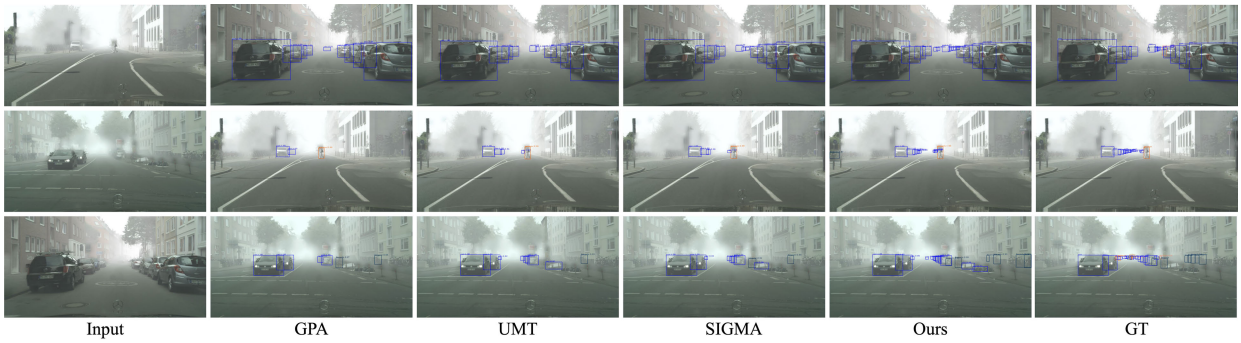

Fig. 7. The qualitative object detection results of different algorithms on the Foggy Cityscapes dataset for Cityscapes→Foggy Cityscapes adaptation. Best viewed in color.

TABLE VII

CROSS-DOMAIN OBJECT DETECTION ON THE FOGGY CITYSCAPES DATASET USING CITYSCAPES→FOGGY CITYSCAPES ADAPTATION. THE PER-CLASS AP (AP_{50}) IS ALSO PROVIDED

Methods	person	rider	car	truck	bus	train	motor	bicycle	mAP_{50} ↑
DA-faster [41]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.3	27.6
SCL [89]	31.6	44.0	44.8	30.4	41.8	40.7	33.6	36.2	37.9
GPA [90]	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
UMT [91]	**56.5**	37.3	48.6	30.4	33.0	**46.7**	**46.8**	34.1	41.7
SIGMA [92]	43.9	49.9	60.6	29.6	50.7	39.0	38.3	42.8	44.3
Ours	52.4	**56.4**	**72.3**	**30.5**	**52.9**	28.8	39.3	**51.4**	**48.0**

TABLE VIII

THE QUANTITATIVE RESULTS BETWEEN DIFFERENT ALGORITHMS ON THE SEEINGTHROUGHFOG DATASET [83]. THE BEST RESULTS ARE IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED

Settings	Dense fog			snow/rain		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Image-only SSD [83]	**87.89**	<u>78.25</u>	<u>74.96</u>	84.33	74.38	67.01
CYCADA [34]	87.24	77.04	73.38	85.56	74.80	67.22
ADDA [33]	<u>87.64</u>	78.12	74.37	84.17	74.25	66.86
DEEP FUSION [83]	86.77	77.28	73.93	**89.25**	**79.09**	**70.51**
Ours+SSD	87.45	**78.68**	**75.21**	<u>87.67</u>	<u>76.03</u>	<u>68.45</u>

the pre-trained mono depth estimation model[4]. The qualitative results are illustrated in Fig. 9. As reported, the image enhancement could lead to more consistent and accurate depth estimation results. Considering the disparity ground truth of the Cityscapes dataset is provided and both Foggy Cityscapes and Rainy Cityscapes are rendered from the Cityscapes, thus the two datasets have the same disparity ground truth as the Cityscapes dataset. Following the evaluation metrics of [13], we report the quantitative results in Table IX. Under the *Oracle* setting, the clear images from the Cityscapes dataset are used for evaluation.

F. Discussions

In this section, we discuss the effectiveness of the mentioned different settings in this paper and the limitation

[4]We use the pre-trained model trained on KITTI dataset from https://github.com/nianticlabs/monodepth2. The model resolution is set to 640 × 192.

Fig. 8. The qualitative semantic segmentation results under various settings on Alderley dataset [23].

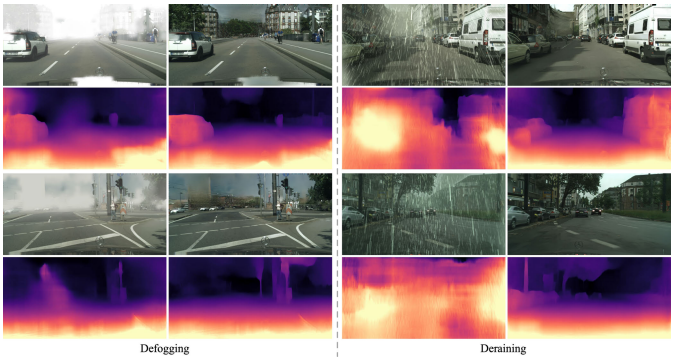

Fig. 9. The qualitative mono depth estimation results under the defogging and deraining settings. We can achieve significant depth estimation performance improvement with visibility enhancement.

TABLE IX
THE QUANTITATIVE DEPTH ESTIMATION RESULTS OF 500 VALIDATION IMAGES ON THE FOGGY CITYSCAPES DATASET UNDER TWO SETTINGS: WITHOUT AND WITH *Enhancement*

Method	Error↓				Accuracy↑		
	RMSE	RMSE(log)	Abs Rel	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
w/o Defogging	14.35	0.444	0.319	4.844	0.448	0.725	0.861
w/ Defogging	**11.50**	**0.336**	**0.239**	**3.064**	**0.567**	**0.837**	**0.933**
w/o Deraining	20.60	0.723	0.732	16.04	0.210	0.409	0.593
w/ Deraining	**12.43**	**0.357**	**0.246**	**3.345**	**0.550**	**0.814**	**0.918**
Oracle	11.45	0.333	0.236	3.055	0.574	0.839	0.934

of the proposed method. For the *Enhancement* setting, the unpaired I2I synthesis algorithms could achieve reasonable image enhancement and when the visibility gap is small (*e.g.*, the daytime↔rainy or daytime↔foggy). However, when the visibility distribution is huge (*e.g.*, daytime↔nighttime or daytime↔rainy night), the unpaired I2I synthesis cannot preserve the meaningful content representations after translation and the visual artifacts will be inevitably introduced. The ineffective translation can even lead to a performance drop. Addressing this, the normal-to-adverse image translation can be adopted to perform data *Augmentation*. The annotations from the source domain are inherited for the generated images and both the original and synthesized images are utilized to retrain visual perception models in the adverse domain. This method could usually achieve more performance gains than the synthesis for enhancement when the distribution shift is huge. For the latter *Two-branch* design, it combines the information from both the original and translated images for feature fusion to achieve a more reliable and robust visual perception.

We did not conduct mono depth estimation experiments under the *Augmentation* setting since the mono depth estimation model is optimized by the geometric correspondences between frames. The current unpaired I2I synthesis is still conducted frame by frame without considering the motions between frames. Besides, extracting geometric correspondences between frames requires a very high-quality image synthesis, which the current proposed unpaired I2I image synthesis cannot meet.

V. CONCLUSION

In this paper, we comprehensively performed the analysis of the potential of choosing synthesis-based data augmentation to achieve robust and accurate visual perception under various adverse conditions. To remedy the noise and uncertainty caused by the unpaired I2I synthesis, we propose a novel and effective two-branch architecture, which utilizes the integration of the raw lossless observations and the synthesized counterparts. We have hierarchically explored the performance improvement and limitations of *Enhancement*, *Augmentation* and *Two-branch* and performed a direct comparison between them. The comprehensive experiments on various datasets have demonstrated the superior performance of the proposed method. We believe our work will help the community understand the effectiveness of synthesis-based data augmentation in practice. Besides, our work can also help drive further advances for more effective and efficient data augmentation.

REFERENCES

[1] J. Levinson et al., "Towards fully autonomous driving: Systems and algorithms," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 163–168.
[2] S. Brandon and S. Michael, "A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia," Univ. Michigan, Ann Arbor, Transp. Res. Inst., Ann Arbor, MI, USA, Tech. Rep. UMTRI-2014-21, 2014.
[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.
[4] M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, *Autonomous Driving: Technical, Legal and Social Aspects*. Cham, Switzerland: Springer, 2016.
[5] M. Hnewa and H. Radha, "Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 53–67, Jan. 2021.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[10] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *IEEE/CVF Int. Conf. Comput. Vis. (CVPR)*, Oct. 2019, pp. 9656–9665.
[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[12] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611.
[13] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837.
[14] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.

[15] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[16] X. Huang et al., "The ApolloScape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1067–10676.

[17] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[18] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[19] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.

[20] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, Jan. 2017.

[21] C. Sakaridis, D. Dai, S. Hecker, and L. V. Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 707–724.

[22] S. Halder, J.-F. Lalonde, and R. D. Charette, "Physics-based rendering for improving robustness to rain," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10202–10211.

[23] M. J. Milford and Gordon. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.

[24] H. Porav, T. Bruls, and P. Newman, "I can see clearly now: Image restoration via de-raining," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7087–7093.

[25] L. Sun, K. Wang, K. Yang, and K. Xiang, "See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion," in *Proc. Artif. Intell. Mach. Learn. Defense Appl.*, Sep. 2019, pp. 77–89.

[26] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 102–118.

[27] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Annu. Conf. Robot Learn.*, 2017, pp. 1–16.

[28] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "VirtualWorlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4340–4349.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[30] S. Zhao et al., "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.

[31] H. Porav, T. Bruls, and P. Newman, "Don't worry about the weather: Unsupervised condition-dependent domain adaptation," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 33–40.

[32] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2100–2110.

[33] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[34] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[35] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool, "Night-to-day image translation for retrieval-based localization," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5958–5964.

[36] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[37] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[38] D. Brüggemann, C. Sakaridis, P. Truong, and L. Van Gool, "Refign: Align and refine for adaptation of semantic segmentation to adverse conditions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3173–3183.

[39] S. Di et al., "Rainy night scene understanding with near scene semantic adaptation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1594–1602, Mar. 2021.

[40] C. Song, J. Wu, L. Zhu, M. Zhang, and H. Ling, "Nighttime road scene parsing by unsupervised domain adaptation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3244–3255, Apr. 2022.

[41] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[42] C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, "GAN-based day-to-night image style transfer for nighttime vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 951–963, Feb. 2021.

[43] Z. Zheng, Y. Wu, X. Han, and J. Shi, "ForkGAN: Seeing into the rainy night," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 155–170.

[44] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1685–1694.

[45] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[46] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3932–3941.

[47] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 718–731.

[48] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 179–196.

[49] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 35–51.

[50] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2020, pp. 36–52.

[51] Z. Zheng, Z. Yu, H. Zheng, Y. Yang, and H. T. Shen, "One-shot image-to-image translation via part-global learning with a multi-adversarial framework," *IEEE Trans. Multimedia*, vol. 24, pp. 480–491, 2022.

[52] Z. Zheng, Y. Bin, X. Lu, Y. Wu, Y. Yang, and H. T. Shen, "Asynchronous generative adversarial network for asymmetric unpaired image-to-image translation," *IEEE Trans. Multimedia*, vol. 25, pp. 2474–2487, Feb. 2022.

[53] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst. (Neurips)*, vol. 33, 2020, pp. 1–13.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[55] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–13.

[56] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[57] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[58] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.

[59] Z. He and L. Zhang, "Multi-adversarial faster-RCNN for unrestricted object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6667–6676.

[60] P. T. Jackson, A. A. Abarghouei, S. Bonner, T. P. Breckon, and B. Obara, "Style augmentation: Data augmentation via style randomization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 83–92.

[61] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," 2017, *arXiv:1702.05538*.

[62] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 694–710.

[63] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9240–9249.

[64] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (Neurips)*, vol. 27, 2014, pp. 1–9.

[65] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

[66] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5967–5976.

[67] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[68] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst. (Neurips)*, 2017, pp. 700–708.

[69] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–18.

[70] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *Proc. Adv. Neural Inf. Process. Syst. (Neurips)*, 2018, pp. 2108–2118.

[71] D. Bhattacharjee, S. Kim, G. Vizier, and M. Salzmann, "DUNIT: Detection-based unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4786–4795.

[72] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[73] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[74] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng, "Discriminative region proposal adversarial networks for high-quality image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2018, pp. 770–785.

[75] Z. Zheng, J. Yang, Z. Yu, Y. Wang, Z. Sun, and B. Zheng, "Not every sample is efficient: Analogical generative adversarial network for unpaired image-to-image translation," *Neural Netw.*, vol. 148, pp. 166–175, Apr. 2022.

[76] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[77] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[78] M.-Y. Liu et al., "Few-shot unsupervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10550–10559.

[79] Z. Zheng, Z. Yu, Y. Wu, H. Zheng, B. Zheng, and M. Lee, "Generative adversarial network with multi-branch discriminator for imbalanced cross-species image-to-image translation," *Neural Netw.*, vol. 141, pp. 355–371, Sep. 2021.

[80] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1312–1318.

[81] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020.

[82] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, "Deep learning-based image recognition for autonomous driving," *IATSS Res.*, vol. 43, no. 4, pp. 244–252, Dec. 2019.

[83] M. Bijelic et al., "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11679–11689.

[84] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[85] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[86] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[87] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[88] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

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

[90] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12352–12361.

[91] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4089–4099.

[92] W. Li, X. Liu, and Y. Yuan, "SIGMA: Semantic-complete graph matching for domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5281–5290.

Ziqiang Zheng received the B.Eng. degree in communication engineering from the Ocean University of China in 2019. His research interests include deep learning and computer vision.

Yujie Cheng received the B.Eng. degree in information engineering from the Qingdao University of Science and Technology in 2020. He is currently pursuing the Graduate degree with the Department of Information Science and Engineering, Ocean University of China. His research interests include deep learning and computer vision.

Zhichao Xin received the B.S. degree from the Department of Electronic Engineering, Dezhou University, China. He is currently pursuing the M.S. degree with the Department of Electronic Engineering, Ocean University of China. His research interests include deep learning, SLAM, and 3D reconstruction.

Zhibin Yu (Member, IEEE) received the B.E. degree from the Harbin Institute of Technology, China, and the M.E. degree in computer engineering and the Ph.D. degree in electrical engineering from Kyungpook National University, South Korea, in 2009 and 2016, respectively.

In 2016, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently an Associate Professor with the College of Electronic Engineering. His current research interests include underwater image processing, image-to-image translation, and generative neural networks.

Bing Zheng (Member, IEEE) received the B.S. degree in electronics and information systems, the M.S. degree in marine physics, and the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 1991, 1995, and 2013, respectively. He is currently a Professor with the College of Electronic Engineering, Ocean University of China. His research interests include ocean optics, underwater imaging, and optical detection.