CompUDA: Compositional Unsupervised Domain Adaptation for Semantic Segmentation under Adverse Conditions

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

***Abstract*— In autonomous driving, performing robust semantic segmentation under adverse weather conditions is a long-standing challenge. Imperfect camera observations under adverse conditions result in images with reduced visibility, which hinders label annotation and semantic scene understanding based on these images. A common solution is to adopt semantic segmentation models trained in a source domain with ground truth labels and perform unsupervised domain adaptation (UDA) from the source domain to an unlabeled target domain that has adverse conditions. Due to imperfect visual observations in the target domain, such adaptation needs special treatment to achieve good performance. In this paper, we propose a new *compositional* unsupervised domain adaptation (CompUDA) method that disentangles the domain gap based on multiple factors including *style*, *visibility*, and *image quality*. The domain gaps caused by these individual factors can then be addressed separately by introducing the intermediate domains. Specifically, 1) to address the style gap, we perform source-to-intermediate domain adaptation and generate pseudo-labels for self-training in the target domain; 2) to address the visibility gap, we perform a geometry-aligned normal-to-adverse image translation and introduce a synthetic domain; 3) finally, to address the image quality gap between the synthetic and target domain, we perform a synthetic-to-real adaptation based on the generated pseudo-labels. Our compositional unsupervised domain adaptation can be used in conjunction with a wide variety of semantic segmentation methods and result in significant performance improvement across datasets. The codes are available at `https://github.com/zhengziqiang/CompUDA`.**

I. Introduction

Semantic scene understanding is an important and long-standing problem in computer vision as it has been widely adopted in applications such as autonomous driving, robot grasping and navigation, and medical analysis. Let us take autonomous driving as the context for this paper. Making semantic scene understanding work robustly for different scenarios is an extremely challenging task. State-of-the-art methods for semantic segmentation are mainly based on supervised learning and thus require a large amount of pixel-level annotations for training. In autonomous driving, these labels are generally acquired under normal conditions such as daytime and clear weather, since images captured under such conditions have the best details for annotation. However, the trained model on these normal conditions may generalize poorly to *adverse* weather conditions, including nighttime, rainy, foggy, and snowy conditions. This is because images acquired under these adverse conditions are imperfect, i.e., the images tend to have significant differences with images under normal conditions such as appearance, visibility, image quality, etc., making it challenging to predict robust and accurate pixel-level semantic outputs.

A common approach to improve semantic scene understanding under adverse conditions is to perform domain adaptation for models trained in the source domain to work in the target domain. Several domain adaptation methods exist, and a common setting is unsupervised domain adaptation, which aims at transferring the shared knowledge from a labeled source domain to a new target domain without labels. In the context of autonomous driving, domain adaptation for the semantic segmentation task has achieved a wide range of success, including synthetic-to-real adaptation, cross-camera adaptation, and cross-city adaptation. However, a limitation is that these methods generally consider domain gaps as a single entity, where image differences between the source domain and the target domain are explained by a sole dominant factor such as time of day, location, visibility, etc. Unfortunately, the real-world domain gap is *compositional*; there is often more than one factor that causes data difference between the source and the target domain, and these factors are mixed in the observations of the target domain. Addressing compositional domain gaps is therefore of great interest in domain adaption for real applications.

Intuitively, we could directly perform the domain adaptation between the source and the target domain, as depicted by the red arrow in Fig. 1. Due to the huge domain shift between the source domain and the target domain due to mixed factors, the domain adaptation performance is limited [1], [2], [3], [4]. Recent work [5], [6], [7] decouples the domain adaptation problem into a source-to-intermediate domain adaptation and an intermediate-to-target domain adaptation. These methods aim to cumulatively adapt style and visibility shift from the source domain to the target domain, as depicted by the blue arrows in Fig. 1. However, it remains challenging to perform domain adaptation between the reference images in the intermediate domain and the target images especially when the target images are collected under some adverse conditions (*e.g.*, nighttime and rainy nighttime). The cumulative UDA algorithms also heavily suffer from such visibility degradation problems.

In this work, we proposed a compositional unsupervised domain adaptation (**CompUDA**) framework, which contains

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

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

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

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

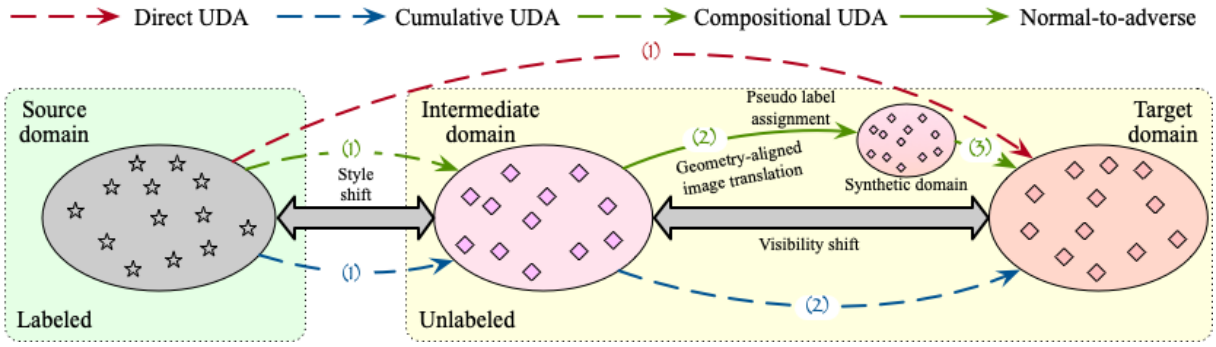

Fig. 1. The comparison between the previous UDA algorithms and our CompUDA. Traditional UDA simply adapts between the source and the target domain (red arrow). Cumulative UDA introduces the intermediate domain (blue arrows) to address the domain gap via a two-stage procedure. Our compositional UDA method (green arrows) further decomposes the intermediate-to-target domain adaptation into smaller steps. We introduce a new synthetic domain in the middle to model the visibility gap by a geometry-aligned image translation step and a synthetic-to-real translation step, in which we utilize the information asymmetry between normal and adverse conditions for better image synthesis. Best viewed in color.

three main procedures as described by the green arrows in Fig. 1: 1) Source-to-intermediate domain adaptation for addressing the style shift; 2) Geometry-aligned image translation for reducing the visibility shift between the reference images in the intermediate domain and the target images; and 3) Synthetic-to-real domain adaptation with pseudo-label assignment. Compared to previous methods, our adaptation for intermediate-to-target domain is more sophisticated. Unlike previous cumulative UDA algorithms, we address the visibility shift by further decomposing it into normal-to-adverse geometry-aligned image translation and synthetic-to-real adaptation. The pseudo-labels generated in the intermediate domain can be used to supervise the synthetic-to-real adaptation since the generated pseudo-labels are inherited for the synthesized counterparts from the reference images. Our performance gain is achieved by more reliable pseudo-labels in the intermediate domain (rather than the target domain in previous direct UDA and cumulative UDA algorithms) and better image synthesis from the normal-to-adverse geometry-aligned image translation. Our compositional domain adaptation method is general and can improve upon different semantic scene understanding methods. The main contributions of this paper are as follows:

- We introduce a novel compositional domain adaptation framework that decouples a general domain gap into individual factors that can be addressed separately.
- We address the visibility shift by introducing a novel intermediate domain known as the synthetic image domain, obtained by a geometry-aligned image translation step and further adapted by a synthetic-to-real translation step.
- We demonstrate the state-of-the-art performance of our CompUDA on different semantic segmentation methods across datasets.

II. Related Work

Domain adaptive semantic segmentation algorithms aim to transfer the learned knowledge in the source domain to the target domain at the pixel level. The Maximum Mean Discrepancy (MMD) [8] algorithms are introduced to learn the domain-invariant representations by minimizing the domain discrepancy. Liu *et al.* [9] utilize the Kullback-Leibler divergence on the mean and variance stored in the batch normalization layer of the model to make the data distributions similar to each other. The adversarial training [10], [11] for domain adaptation is to model the domain style shift in a min-max manner with a domain classifier. Hoffman *et al.* [10] first applied the adversarial training for domain adaptive semantic segmentation and designed a specific category adaptation strategy by transferring the label statistics of the source domain to the target domain. The self-training strategy [12] first generates the pseudo labels in the target domain considering the structural consistency across domains and utilizes such pseudo labels for retraining the segmentation model in the target domain. Chen *et al.* [12] proposed the max square loss based on the gradient to alleviate the imbalanced class distribution. The knowledge distillation [13], [14] is to transfer the learned knowledge from a teacher network to a student network by minimizing the divergence between the predicted distribution of these two networks. DAFormer [13] adopted the SegFormer [15] into the domain adaptation and achieve a significant performance improvement compared with the non-transformer algorithms. The recent state-of-the-art HRDA [14] performs the high-resolution domain adaptation based on the multi-resolution training and fusion strategy.

Domain adaptive semantic segmentation under adverse conditions aims to transfer the semantic segmentation model from the normal condition to various adverse conditions. MGADA [6] adopted an intermediate domain (twilight domain) to gradually reduce the distribution discrepancy. DANNet [2] uses a style translation network to transform different domains into the same style. HeatNet [16] additionally uses thermal data that is not sensitive to illumination. Gao *et al.* [17] leverage the domain shift and regard the cross-domain correlation as the concrete representation of the domain shift to conduct domain adaptation. The recent Refign [18] adopts the pre-trained alignment module to warp the reference image to refine the generated pseudo label in the target domains. However, the geometry alignment requires training with large-scale additional datasets as well as

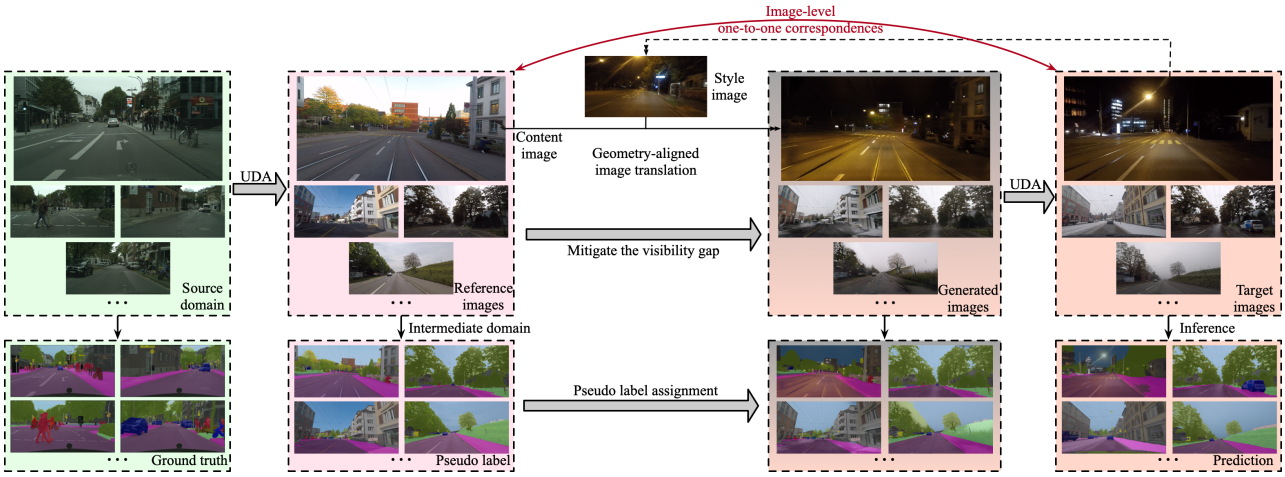

Fig. 2. The framework overview of the proposed method. The reference images collected under the normal condition (daytime) are regarded from the *intermediate* domain. We first perform the unsupervised domain adaptation (UDA) between the source domain and the intermediate domain and generated the pseudo labels for the daytime images. We assume the availability of image-level one-to-one correspondences (geometry alignment) between the daytime images and the low visibility images captured under various adverse conditions. We propose to use a geometry-aligned image translation to transfer the daytime images from the intermediate domain to a synthetic domain for reducing the visibility gap. We then conduct a final synthetic-to-real adaptation between the synthesized images and the target images.

running structure-from-motion which is expensive to achieve. In contrast, our proposed method does not require any pre-trained module or additional training data.

III. Our Method

A. Overview

We first provide the preliminaries for cross-domain semantic segmentation. N_s labeled source images $\{x_s^i\}_{i=1}^{N_s}$ from the source domain $\mathcal{S}$ are provided with the corresponding dense pixel-level annotations $\{y_i^s\}_{1=1}^{N_s}$. We assume the availability of paired reference images $\{x_r^i\}_{i=1}^{N_r}$ (from the intermediate domain $\mathcal{I}$) and target images $\{x_t^i\}_{i=1}^{N_t}$ (from the target domain $\mathcal{T}$) with one-to-one geometric correspondences (x_r^i, x_t^i), where $N_r = N_t$ indicating the number of reference images and target images. It is worth noting that both reference images and target images are unlabelled. Our goal is to transfer the segmentation knowledge from the source domain $\mathcal{S}$ to the target domain $\mathcal{T}$. We model the domain gap between the source images and the target images by a composition of individual factors including **style** and **visibility** shifts, where we further decompose the visibility shift into a geometry-aligned image translation and a synthetic-to-real translation characterized by a new synthetic domain. Our framework is illustrated in Fig. 1 and Fig. 2, which include *source-to-intermediate* adaptation, *geometry-aligned image translation*, and *synthetic-to-real* adaptation as the main steps along with a *pseudo-label assignment* step.

B. Source-to-Intermediate Adaptation

We first conduct the UDA between the source images and the reference images. Since only the labels for source images are available, the supervised cross-entropy loss $\mathcal{L}_{ce}^{\mathcal{S}}$ can only be calculated based on the predictions and the source labels as follows:

$$\mathcal{L}_{ce}^{\mathcal{S}} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} y_{ijc}^{s} \log \hat{y}_{ijc}^{s}, \tag{1}$$

where H and W are the image height and width of source images respectively. C indicates the pre-defined number of the total semantic categories. y_{ijc}^{s} is the given source label and $\hat{y}_{ijc}^{s}$ indicates the predicted class distribution based on the source image. We adopt the self-training strategy to generate the pseudo-labels and the model is iteratively adapted to the intermediate domain by optimizing it with the generated pseudo labels as follows:

$$\mathcal{L}_{ce}^{\mathcal{I}} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} (y_{ijc}^{r}|x_r) \log \hat{y}_{ijc}^{r}, \tag{2}$$

where y_{ijc}^{r} is the pseudo-label generated based on the reference image x_r from the intermediate domain, and $\hat{y}_{ijc}^{r}$ is the predicted label. Considering that both the source images and the reference images are collected under the normal condition, with only the style shift, the model could generate more reliable and accurate pseudo-labels in the intermediate domain compared with the target domain.

C. Geometry-aligned Image Translation

We proposed to conduct effective and reliable image synthesis for reducing the visibility gap between the normal condition and various adverse weather conditions based on reference-guided image synthesis [19], [20]. In this section, we perform the geometry-aligned image translation, which aims at generating target-like images with the same source content, which can reduce the visibility shift. Given the geometry-aligned pair (x_r, x_t), we fed x_r into a content network to extract the content representations and fed x_t to a style network to extract its visibility representations. We adopt FAdaIN for style transfer from x_t to x_r as follows:

$$\text{FAdaIN}(z, f_{x_t}) = \sigma(f_{x_t}) \left(\frac{z - \mu(z)}{\sigma(z)} \right) + \mu(f_{x_t}), \tag{3}$$

where $\mu(z)$ and $\sigma(z)$ are the mean and standard deviation of the fused feature representation z for geometry-aligned

image synthesis, and f_{x_t} is the style feature representation extracted from the target image x_t. To stabilize the entire training procedure, we use the hinge-based adversarial loss and define the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$ as:

$$\mathcal{L}_G = -\mathbb{E}[D(G(x_r, x_t))] + \mathcal{L}_{FM}(G(x_r, x_t), x_r), \quad (4)$$

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}[\min(-1 + D(x_t), 0)] \\ &-\mathbb{E}[\min(-1 - D(G(x_r, x_t)), 0)],\end{aligned} \quad (5)$$

where we adopt the feature matching loss $\mathcal{L}_{FM}$ designed in Pix2pixHD [21] to match the intermediate feature representations between the generated images and the original reference images at different layers of a multi-scale discriminator [19].

Pseudo-label assignment. After optimizing the geometry-aligned image translation model, our objective is to generate pseudo-labels for the synthesized images $\tilde{x}_r = G(x_r, x_t)$ in the target domain based on the aligned reference-target image pairs. Particularly, we assign the pseudo-labels of the reference images to $\tilde{x}_r$ for constructing the image-label pairs $(\tilde{x}_r, y^r)$, where y^r is the pseudo-label generated based on the original reference image x_r at the former source-to-intermediate domain adaptation.

D. Synthetic-to-Real Adaptation

Finally, we perform the synthetic-to-real adaptation between the synthetic domain and the target domain to reduce the image quality gap. Similarly, we compute the supervised cross-entropy loss in the synthetic image domain as follows:

$$\mathcal{L}_{ce}^{\tilde{\mathcal{I}}} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} y_{ijc}^r \log(\hat{y}_{ijc}^r | G(x_r, x_t)), \quad (6)$$

where the $\hat{y}_{ijc}^r$ is the predicted class distribution based on the synthesized image $G(x_r, x_t)$. Importantly, we do not address the visibility shift in the proposed framework, and in contrast, we aim to reduce the image quality gap by generating the pseudo-labels in the target domain and constructing the supervision as:

$$\mathcal{L}_{ce}^{\mathcal{T}} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} (y_{ijc}^t | x_t) \log \hat{y}_{ijc}^t, \quad (7)$$

where we could obtain more reliable domain adaptation since both the synthesized images and the target images are with similar appearance representations and the image quality gap is smaller than the visibility shift. Note that the two UDA procedures (source-to-intermediate and synthetic-to-real) are optimized separately. The UDA and geometry-alignment image translation components in our method could also be replaced with other alternative counterparts for more effective domain adaptation.

IV. Experiments

A. Implementation Details and Datasets

Implementation details. For all the domain adaptive semantic segmentation under various adverse conditions, the source domain is set to clear Cityscapes dataset. As for the target domain data, we adopted the ACDC dataset [22], Dark Zurich dataset [23], the nighttime images in the BDD100K dataset [24], Nighttime Driving [25] and Alderley dataset [26]. To demonstrate the flexibility of the proposed method, we combine our method with three state-of-the-art UDA methods: DACS [1] (DeepLabv2 [27] based); DAFormer [13] (SegFormer based); HRDA [14] (DAFormer based). We have achieved significant improvement among different baselines. The image resolution is set to 2048×1024 (512×256 for Cityscapes-to-Alderley adaptation) to perform high-resolution image generation.

Datasets. **Cityscapes dataset** [28] is a real-world dataset composed of street view images captured in 50 different cities. Its data split includes 2,975 training images and 500 validation images. All the images from this dataset are regarded as source images and labeled with dense pixel-wise category annotations. **ACDC dataset** [22] contains four adverse-condition categories (fog, rain, snow, and nighttime) with pixel-level annotations. Each of them contains 1,000 images and is split into the training set, validation set, and testing set for roughly 4:1:5 proportion. The annotations for the test set are withheld for online evaluation. **Dark Zurich** dataset [23] is captured in Zurich, with 3,041 daytime, 2,920 twilight, and 2,416 nighttime images for training, which are all unlabeled with a resolution of $1,920 \times 1,080$. Each nighttime image has a corresponding daytime image as the auxiliary image, which constitutes an image pair for geometry-aligned image synthesis in our proposed framework. We use the 2,416 night-day image pairs in our training process. The Dark Zurich also contains 201 manually annotated nighttime images, of which 151 (Dark Zurich-test) are used for testing and 50 (Dark Zurich val) are used for validation. Note that the evaluation of the Dark Zurich test only serves as an online benchmark, and its ground truth is not publicly available. The testing images from the Foggy Zurich and BDD100K datasets are adopted for testing the generalization ability of the optimized segmentation model. **Alderley** dataset [26] is a vision dataset gathered from a car driven around Alderley, Queensland in two different conditions for the same route: one on a sunny day and one during a rainy night. The day-night correspondences are provided based on the GPS alignment. Due to that there are no manually annotated dense pixel-level semantic segmentation annotations, we only provide qualitative results.

B. Comparisons with SOTAs

ACDC. We compare our proposed method with some existing state-of-the-art methods, including DMAda [25], GCMA [23], MGCDA [23], DANNet [2], and several other domain adaptation approaches [13], [14], [18] on Dark Zurich-test and ACDC dataset. The MGCDA, GCMA, DMAda, and DANNet adopt the RefineNet [29] as the baseline, while other methods use the Deeplab-v2 [27]. In our proposed method, we choose SegFormer [15] as the network backbone considering its powerful ability to extract the feature representations. To make a fair comparison, we also report the experimental results based on RefineNet and Deeplab-v2. The comparisons

TABLE I

Comparison with previous UDA methods on the Cityscapes → ACDC domain adaptation.

Methods	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU↑
MGCDA (RefineNet) [6]	73.4	28.7	69.9	19.3	26.3	36.8	53.0	53.3	75.4	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DANNet (PSPNet) [2]	84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	86.6	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
DANIA (PSPNet) [3]	88.4	60.6	81.1	37.1	32.8	28.4	43.2	42.6	77.7	50.5	90.5	51.5	31.1	76.0	37.4	44.9	64.0	31.8	46.3	53.5
DACS [1]	58.5	34.7	76.4	20.9	22.6	31.7	32.7	46.8	58.7	39.0	36.3	43.7	20.5	72.3	39.6	34.8	51.1	24.6	38.2	41.2
CompUDA (DACS)	52.4	54.5	75.6	30.6	26.8	35.6	44.7	47.8	74.5	40.5	39.1	45.1	20.6	76.3	47.2	40.5	64.9	36.2	40.1	47.0
DAFormer [13]	56.9	45.4	84.7	44.7	35.1	48.6	44.8	57.4	69.5	52.9	45.8	57.1	28.2	82.8	57.2	63.9	84.0	40.2	50.5	55.3
DAFormer†	57.4	48.7	85.1	43.5	38.7	50.1	46.1	58.1	67.3	52.8	49.1	56.1	28.1	84.2	59.1	67.1	81.5	43.5	50.5	56.2
CompUDA (DAFormer)	92.5	**71.6**	87.2	45.3	39.8	54.2	**70.2**	68.2	**86.4**	50.8	**94.8**	65.2	45.5	87.2	60.9	69.9	84.5	50.6	59.9	67.6
HRDA [14]	88.3	57.9	88.1	55.2	36.7	**56.3**	62.9	65.3	68.8	57.7	85.9	**68.9**	45.7	88.5	**76.4**	**82.4**	87.7	52.7	60.4	68.0
HRDA†	93.0	73.5	87.9	50.4	**42.7**	55.6	71.1	**68.7**	85.9	51.0	94.3	67.5	45.5	87.4	63.7	69.0	79.7	50.7	60.9	68.3
CompUDA (HRDA)	**92.7**	71.5	**89.5**	**61.6**	39.8	51.0	72.0	67.2	82.8	**58.7**	92.9	67.0	**46.4**	**89.3**	75.3	81.2	**88.7**	**56.3**	**62.4**	**71.1**

Fig. 3. Qualitative geometry-aligned image translation results under four different weather conditions on the ACDC dataset. We also provide the reference images with geometry correspondences with the input images for better comparison.

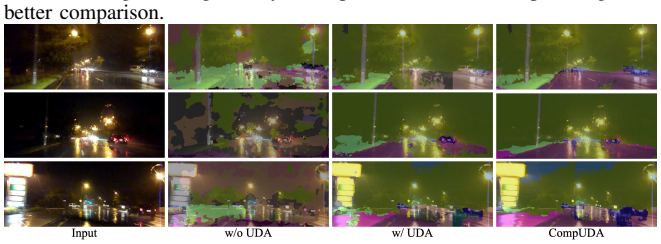

Fig. 4. The qualitative semantic segmentation results generated by different UDA algorithms on the Alderley dataset.

with the recent DAFormer [13], SePiCo [30] and HRDA [14] are also reported. We present comparisons to several state-of-the-art methods on the ACDC test set in Table I. Applying our CompUDA based on DAFormer results in a mIoU of 71.1%, a new state-of-the-art in domain adaptation from Cityscapes-to-ACDC adaptation. Besides, we also report the experimental results of performing the cumulative UDA based on the different UDA algorithms (DAFormer and HRDA), denoted as DAFormer† and HRDA†. We perform domain adaptation twice: source-to-intermediate adaptation and intermediate-to-target adaptation. The generated pseudo labels by the first stage are regarded as the labels for full supervision. As reported, DAFormer† and HRDA† could only achieve marginal improvement compared with the vanilla version since the two methods still suffer from the visibility shift. The results of using various segmentation backbones as reported in Table I have also shown the flexibility of our CompUDA.

We provide qualitative image synthesis results in Fig. 3 to demonstrate that CompUDA could effectively reduce the visibility shift on ACDC dataset [22]. The image synthesis results under four weather conditions: *night*, *snow*, *foggy*, and *rain* are included. Besides, we have also provided the corresponding target images with geometry correspondences for better illustration. As illustrated, the proposed method could effectively simulate the visibility shift to various weather conditions and generate high naturalness images with a small image quality gap with the real target images.

Dark Zurich. We conduct experiments on the Dark Zurich-test benchmark and we compare the recent works in Table II. Following previous works, the trained Dark Zurich models are also tested for evaluating the generalization on Nighttime Driving [25] and BDD100k-night [24] in Table III. The proposed CompUDA definitely achieved the highest 62.9% mIoU, demonstrating a superior ability of our method over the existing algorithms.

Alderley. Similarly, we perform experiments under a more challenging rainy nighttime condition. The light reflection and the raindrops lead to a huge challenge to accurately identify objects. Since there is no ground truth provided on the Alderley dataset, we only provide the qualitative results comparison of different methods in Fig. 4. "w/o UDA" indicates directly performing the semantic segmentation based on a pre-trained segmentation model on the Cityscapes dataset. "w/ UDA" indicates the setting that we perform the UDA (HRDA) between the source images (Cityscapes dataset) and the target rainy nighttime images (Alderley dataset). Compared with these two settings, our CompUDA could achieve better domain adaptive semantic segmentation performance by conducting a compositional unsupervised domain adaptation.

C. Ablation Studies

To better dissect the contribution of each part of the proposed method, we conduct the ablation studies on Cityscapes→ACDC (nighttime) adaptation for better comparison since there is a large visibility shift between the daytime images and the nighttime images. All the experimental results are reported in Table IV. Directly performing **source-to-target** (also daytime-to-nighttime) adaptation leads to 55.2% mIoU. By introducing the daytime reference images from the intermediate domain, the **cumulative UDA** could achieve marginal performance improvement. We also conduct the **source-to-intermediate** domain adaptation to obtain an adaptive segmentation model on the daytime image domain on the ACDC dataset. The **adverse-to-normal** (also nighttime-

TABLE II

Comparison with previous UDA methods on the Cityscapes → Dark Zurich-test set domain adaptation. † indicates to perform the UDA cumulatively. We perform the UDA algorithms from source-to-intermediate and intermediate-to-target adaptations.

Methods	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU↑
DMAda [25]	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
GCMA [23]	81.7	46.9	58.8	22.0	20.0	41.2	40.5	41.6	64.8	31.0	32.1	53.5	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA [6]	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	53.5	74.7	66.0	0.0	37.5	29.1	22.7	42.5
DANNet (RefineNet) [6]	90.0	54.0	74.8	41.0	21.1	25.0	26.8	30.2	72.0	26.2	84.0	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
CCDistill [17]	89.6	58.1	70.6	36.6	22.5	33.0	27.0	30.5	68.3	33.0	80.9	42.3	40.1	69.4	58.1	0.1	72.6	47.7	21.3	47.5
SePiCo [30]	93.2	68.1	73.7	32.8	16.3	54.6	49.5	48.1	74.2	31.0	86.3	57.9	50.9	82.4	52.2	1.3	83.8	43.9	29.8	54.2
DAFormer [13]	92.0	63.0	67.2	28.9	13.1	44.0	42.0	42.3	70.7	28.2	83.6	51.1	39.1	76.4	31.7	0.0	78.3	43.9	26.5	48.5
DAFormer†	94.0	69.1	70.4	35.1	19.4	58.5	52.4	39.4	67.4	17.5	85.6	56.2	43.6	79.5	40.3	1.7	80.6	44.1	30.1	51.8
CompUDA (DAFormer)	**95.7**	**77.4**	**83.6**	**50.0**	34.2	**62.5**	**62.2**	**70.0**	**81.1**	16.7	**91.5**	**67.3**	60.0	**88.1**	5.5	**32.1**	90.8	55.7	41.4	61.4
HRDA [14]	90.4	56.3	72.0	39.5	19.5	57.8	52.7	43.1	59.3	29.1	70.5	60.0	58.6	84.0	75.5	11.2	90.5	51.6	40.9	55.9
HRDA†	91.4	57.4	70.6	44.2	23.6	59.4	55.1	40.3	56.7	34.5	73.1	63.5	**60.4**	80.5	78.3	15.2	90.9	47.8	**43.2**	57.2
CompUDA (HRDA)	94.3	72.2	79.1	47.0	**35.0**	57.6	57.0	59.5	72.8	**40.2**	88.2	62.9	**60.4**	85.7	**87.6**	0.2	**92.3**	**61.5**	41.3	**62.9**

TABLE III

Trained models are tested for generalization on the Nighttime Driving and BDD100k-night test sets.

Method	mIoU ↑	
	Nighttime Driving	BDD100k-night
DMAda (RefineNet) [25]	36.1	28.3
GCMA (RefineNet) [23]	45.6	33.2
MGCDA (RefineNet) [6]	49.4	34.9
CDAda (RefineNet) [31]	50.9	33.8
DANNet (PSPNet) [6]	47.7	28.0
DANIA (PSPNet) [3]	48.4	27.0
CCDistill (RefineNet) [17]	46.2	33.0
SePiCo (DAFormer) [30]	57.1	36.9
DAFormer [13]	51.8	34.2
Ours (DAFormer)	**59.1**	**38.4**

to-daytime) image synthesis is then conducted to transfer the testing nighttime images to the daytime domain for evaluation. However, we observe a significant performance drop since it is a very challenging task to synthesize high-quality daytime image outputs from degraded nighttime images. Furthermore, without a compositional manner, we directly perform the **source-to-target** image synthesis and also the **synthetic-to-real** domain adaptation to bridge the final domain gap. Without the geometry correspondence, the image synthesis could not introduce a significant performance gain. In contrast, our CompUDA separates the mixed domain gap into three separate style, visibility, and image quality factors, leading to the best domain adaptive semantic segmentation performance among all the settings.

Does visibility matter? We explored different scales of improvement under different conditions, *e.g.,* rainy and nighttime. The rainy images have the best visibility among the four conditions on ACDC dataset while in contrast, the nighttime images have poor visibility. We compare the proposed method with the HRDA on the two settings in Table V. HRDA cannot achieve a satisfactory performance under the nighttime setting and our method has gained a large gain over HRDA. Meanwhile, only marginal improvement has been achieved in the rainy setting.

Comparison with Refign. The recent work Refign [18] could achieve better results (72.1% vs. 71.1% mIoU on ACDC dataset and 63.9% vs. 62.9% mIoU on Dark Zurich dataset) than the proposed CompUDA based on a pre-trained geometry alignment module on additional large-scale data. Besides, the geometry alignment module is also optimized simultaneously with the semantic segmentation module. Refign highly depends on geometric matching and requires additional geometric constraints from extra annotation. However, our CompUDA only requires image-level correspondence, which is easy to obtain. Our method is favorable in that it is task-agnostic and does not require additional labels for training. The proposed method could also be combined with other domain adaptation algorithms seamlessly.

V. Conclusion

In this paper, we propose a compositional unsupervised domain adaptation framework to address the problem of semantic segmentation under adverse conditions. We validate the effectiveness of properly handling two kinds of domain shifts, *i.e.* style and visibility difference, where the visibility shift is further decomposed and addressed by a geometry-aligned image translation and a synthetic-to-real adaptation via a new synthetic domain. Experimental results confirm the effectiveness of our proposed method across datasets.

Despite our state-of-the-art performance, our method is not without limitations. Since the distillation is based on the domain shift between the source and target domain, it cannot always be effective enough for all adverse conditions, which will be further explored in our future work. Directly performing semantic segmentation on the generated target-like images cannot lead to a large performance gain since the image synthesis might introduce some unsatisfactory visual artifacts. Addressing this problem is left to our future work.

References

[1] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR)*, pp. 1379–1389, 2021.

[2] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15769–15778, 2021.

[3] X. Wu, Z. Wu, L. Ju, and S. Wang, "A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 01, pp. 1–1, 2021.

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

[5] D. Dai, C. Sakaridis, S. Hecker, and L. Van Gool, "Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 5, pp. 1182–1204, 2020.

TABLE IV

Comparison with previous UDA methods on the Cityscapes → ACDC (nighttime) test-set domain adaptation.

Methods	Settings	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU↑
HRDA	Source-to-target adaptation	89.6	54.9	76.5	43.1	18.8	53.5	49.7	52.2	47.3	46.6	27.7	65.1	35.1	78.4	58.1	**76.6**	80.4	43.4	**52.3**	55.2
HRDA†	Cumulative UDA	94.3	71.6	72.7	37.4	23.7	47.0	50.0	48.5	57.4	41.9	61.4	57.1	45.7	77.0	65.3	61.3	79.0	34.6	44.3	56.3
HRDA	Source-to-Intermediate adaptation + Adverse-to-normal synthesis	85.6	47.6	63.2	35.3	10.5	41.2	45.1	43.6	43.1	20.4	41.5	17.6	**59.7**	50.3	60.5	65.1	64.7	35.6	40.1	45.8
HRDA	Source-to-target synthesis + Synthetic-to-real adaptation	94.8	75.0	78.5	45.7	20.6	50.5	53.2	49.7	66.8	45.7	72.7	64.6	40.8	80.8	**65.3**	54.1	84.0	47.4	47.6	57.3
CompUDA	Proposed	**96.0**	**79.6**	**83.4**	**50.2**	**25.3**	**56.3**	**61.0**	**60.1**	**71.3**	**47.5**	**81.7**	**66.1**	41.5	**83.0**	10.3	67.2	**87.8**	**55.9**	50.2	**61.8**

TABLE V

Comparison with previous UDA methods on the Cityscapes → ACDC test-set domain adaptation under various weather conditions.

Methods	Settings	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU↑
HRDA	Rain	74.4	59.5	91.8	65.2	40.9	59.0	73.5	64.6	90.0	44.6	88.6	68.7	30.8	87.6	72.5	89.6	88.9	56.2	63.4	68.9
CompUDA		91.6	73.8	93.2	55.8	47.5	63.6	75.6	76.4	94.1	44.8	98.6	69.7	36.2	90.3	68.7	90.5	87.2	55.6	65.5	72.6$_{3.7\uparrow}$
HRDA	Fog	91.9	73.4	86.1	48.9	28.6	41.7	60.3	68.2	84.4	66.5	97.0	55.6	39.9	79.9	64.8	33.8	83.6	44.3	35.6	62.3
CompUDA		93.3	74.5	86.0	50.3	31.0	47.1	63.2	68.8	80.8	57.9	93.7	52.2	58.5	84.7	66.5	83.6	84.3	59.4	41.5	67.2$_{4.9\uparrow}$
HRDA	Snow	35.4	45.2	82.5	50.9	44.6	57.3	77.4	61.9	88.1	1.7	44.3	75.3	54.9	83.7	67.3	76.6	86.7	43.6	66.1	60.2
CompUDA		89.2	63.4	88.0	34.3	48.6	59.6	78.1	74.9	92.0	18.1	97.3	76.6	51.0	90.4	63.0	68.7	89.5	43.6	71.1	68.3$_{8.1\uparrow}$
HRDA	Night	89.6	54.9	76.5	43.1	18.8	53.5	49.7	52.2	47.3	46.6	27.7	65.1	35.1	78.4	58.1	76.6	80.4	43.4	52.3	55.2
CompUDA		96.0	79.6	83.4	50.2	25.3	56.3	61.0	60.1	71.3	47.5	81.7	66.1	41.5	83.0	10.3	67.2	87.8	55.9	50.2	61.8$_{6.6\uparrow}$

[6] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[7] X. Ma, Z. Wang, Y. Zhan, Y. Zheng, Z. Wang, D. Dai, and C.-W. Lin, "Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18922–18931, 2022.

[8] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," *Advances in neural information processing systems*, vol. 29, 2016.

[9] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, 2021.

[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning (ICML)*, pp. 1989–1998, Pmlr, 2018.

[11] L. Chen, H. Chen, Z. Wei, X. Jin, X. Tan, Y. Jin, and E. Chen, "Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7181–7190, 2022.

[12] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2090–2099, 2019.

[13] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9924–9935, 2022.

[14] L. Hoyer, D. Dai, and L. Van Gool, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," *European conference on computer vision (ECCV)*, 2022.

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

[16] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8461–8468, IEEE, 2020.

[17] H. Gao, J. Guo, G. Wang, and Q. Zhang, "Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9913–9923, 2022.

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

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

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

[21] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, 2018.

[22] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10765–10775, 2021.

[23] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7374–7383, 2019.

[24] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, vol. 2, no. 5, p. 6, 2018.

[25] D. Dai and L. Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3819–3824, IEEE, 2018.

[26] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1643–1649, IEEE, 2012.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2017.

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[29] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1925–1934, 2017.

[30] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *arXiv preprint arXiv:2204.08808*, 2022.

[31] Q. Xu, Y. Ma, J. Wu, C. Long, and X. Huang, "Cdada: A curriculum domain adaptation for nighttime semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2962–2971, 2021.