WeatherDG: LLM-assisted Procedural Weather Generation for Domain-Generalized Semantic Segmentation

Chenghao Qian*, Yuhu Guo*, Yuhong Mo, Wenjing Li

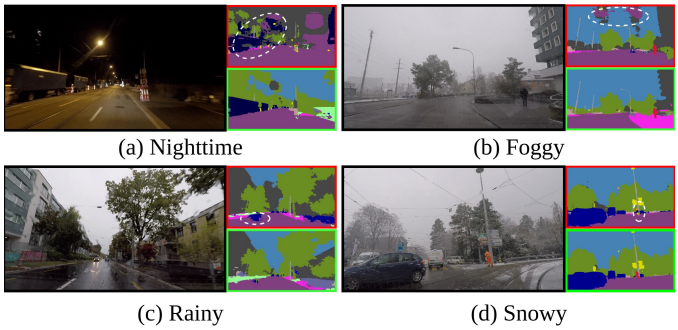
***Abstract*—In this work, we propose a novel approach, namely WeatherDG, that can generate realistic, weather-diverse, and driving-screen images based on the cooperation of two foundation models, i.e, Stable Diffusion (SD) and Large Language Model (LLM). Specifically, we first fine-tune the SD with source data, aligning the content and layout of generated samples with real-world driving scenarios. Then, we propose a procedural prompt generation method based on LLM, which can enrich scenario descriptions and help SD automatically generate more diverse, detailed images. In addition, we introduce a balanced generation strategy, which encourages the SD to generate high-quality objects of tailed classes under various weather conditions, such as riders and motorcycles. This segmentation-model-agnostic method can improve the generalization ability of existing models by additionally adapting them with the generated synthetic data. Experiments on three challenging datasets show that our method can significantly improve the segmentation performance of different state-of-the-art models on target domains. Notably, in the setting of "Cityscapes to ACDC", our method improves the baseline HRDA by 13.9% in mIoU. See the project page for more results:weatherDG.github.io.**

I. Introduction

Semantic segmentation is a fundamental task in autonomous driving. Despite significant achievements in this field, existing models still face serious challenges when deploying in unseen domains due to the well-known domain shift problem [1]. In addition, this issue will be more serious when the unseen domains are with adverse weather conditions [2], [3], such as foggy, rainy, snowy, and nighttime scenarios [4].

One naive way to solve the above problem is collecting more diverse training data. However, labelling segmentation is a time-consuming process, as we need to annotate every pixel in an image. Hence, domain generalization becomes popular in solving the domain shift problem [5], in which the goal is to train a model that can generalize to unseen domains using only the given source data. Existing domain generalization methods can be generally divided into two categories: normalization [6], [7] and data augmentation [8], [9]. In this paper, we focus on the latter category, which is more flexible to different model structures and can be easily integrated with the former techniques.

Data augmentation for domain generalization aims to generate new, diverse and realistic synthetic images for augmenting the training data. Previous methods commonly adopt simulators [11]–[13] or image translation models [14], [15] to generate new samples. Despite their effectiveness, these methods still suffer from diversity and authenticity, particularly in generating samples with adverse weather conditions (shown in Figure 2). In recent years, Stable Diffusion (SD) [16] has shown its strong ability to generate realistic, diverse, and high-quality images. This inspires us to leverage SD to solve the drawbacks of previous data augmentation methods in domain generalization. However, directly applying SD in our task will lead to a critical issue: the styles and layouts are very different from the driving-screen samples (see Figure 2b)). Hence, training with such synthetic samples will hamper the performance of the model on unseen domains. This problem is mainly caused by that the SD was trained with various types of samples, such as natural images and artistic images, instead of specifically with the driving-screen samples. As such, the SD cannot well generate on-the-hand driving-screen-aware samples without a detailed and specific guide.



(a) Nighttime (b) Foggy

(c) Rainy (d) Snowy

Fig. 1: **Visualization of domain-generalized semantic segmentation results: MIC [10] (red box) vs. WeatherDG (green box).** The tested images include foggy, nighttime, snowy, and rainy scenarios.

To solve the above drawback, we propose a novel data augmentation approach named WeatherDG to generate realistic, weather-diverse, and driving-screen images based on Stable Diffusion (SD) and Large Language Model (LLM). Our method is composed of three steps. *Step-I: SD Fine-tuning*. We fine-tune the SD with source data. This enables us to align the content and layout of generated samples with real-world driving scenarios. *Step-II: Procedural Prompt Generation*. We propose a prompt generation method based on LLM, where we leverage the LLM agents to procedurally enrich scenario descriptions (prompt). The generated prompt can help SD to automatically generate more diverse, detailed images. In addition, we introduce a balanced generation strategy to enrich

C. Qian, and W. Li are with the Transport Studies at Institute at University of Leeds, UK

Y. Guo and Y. Mo are with Department of Electrical and Computer Engineering, Carnegie Mellon University, USA

* denotes equal contribution

tailed classes. *Step-III: Sample Generation and model training.* Given the fine-tuned SD and the generated prompts, we can then generate new, diverse samples for model training. The generated samples are used to train the model together with the source data. In sum, our contributions are threefold:

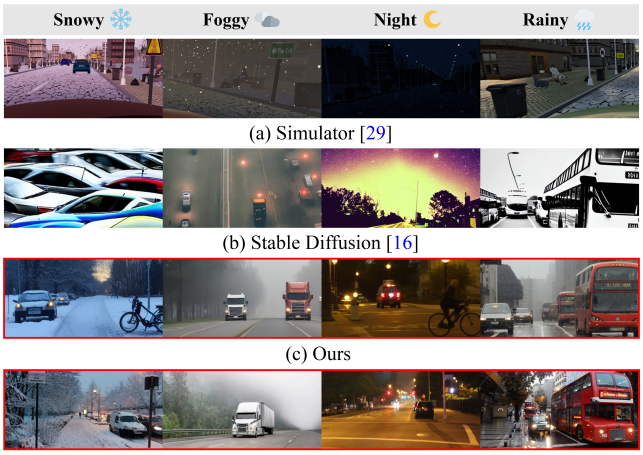
- We propose a novel data augmentation framework based on SD and LLM for domain generalization in adverse weather conditions. Our method can generate realistic, diverse samples for improving the model's generalization ability in unseen domains.
- We propose two novel strategies for prompt generation and sample generation, which encourage SD to generate diverse and driving-screen samples that are beneficial to our segmentation task.
- Our method is segmentation-model-agnostic. Experiments on three challenging datasets demonstrate that our method can consistently improve the performance of state-of-the-art methods.

II. Related Work

Domain Generalization for Semantic Segmentation (DGSS). DGSS aims to train deep neural networks that perform well on semantic segmentation tasks across multiple unseen domains. Existing DGSS methods [5], [17], [18] attempt to address the domain gap problem through two main approaches: normalization and data augmentation. Normalization-based methods [6], [7] train by normalizing the mean and standard deviation of source features or whitening the covariance of these features. Data augmentation-based method [5] transform source images into randomly stylized versions, guiding the model to capture domain-invariant shape features as texture cues are replaced with random styles [19]. For instance, SHADE [20] creates new styles derived from the foundational styles of the source domain, while MoDify [5] utilizes difficulty-aware photometric augmentations.

With the advent of generative diffusion models, several studies [8], [9] have proposed content augmentation to enhance generalization. However, these approaches often either lack realism or fail to adequately consider variations in weather and lighting conditions.

Unsupervised Domain Adaptation (UDA). UDA aims to boosts model performance on domain-specific data without needing labeled examples. Existing UDA techniques can be categorized into three main approaches: discrepancy minimization, adversarial training, and self-training. Discrepancy minimization reduces the differences between domains by using statistical distance functions [21]. Adversarial training involves a domain discriminator within a GAN framework to encourage domain-invariant input [22], feature [23] or output [24]. Self-training generates pseudo-labels for the target domain based on predictions made using confidence thresholds [25] or pseudo-label prototypes [26]. Recently, DATUM [27] introduces a one-shot domain adaptation method that generates a dataset using a single image from the target domain and pairs it with unsupervised domain adaptation training methods to bridge the sim-to-real gap. Futher, PODA [28] leveraged the capabilities of the CLIP model to enable zero-shot domain adaptation using prompts.



(a) Simulator [29]

(b) Stable Diffusion [16]

(c) Ours

(d) Real-world Images

Fig. 2: **Comparison between synthetic and real-world images under adverse weather conditions.** The results reveal that images generated by (a) driving simulator [29] lack intricate details and natural lighting, whereas (b) Stable Diffusion [16] typically produces images with an artistic flair. In contrast, (c) our method produces the most realistic images, closely resembling (d) diverse real-world driving scenes.

Text-based Image Synthesis. The current text-to-image task is predominantly driven by diffusion-based and LLM-oriented methods. Diffusion models, a breakthrough in producing photorealistic images, prompting studies to explore their use in enriching source domain datasets and improving semantic segmentation. For example, DIDEX [8] employs ControlNet [30] to convert synthetic images into real-world styles. Nevertheless, this method often lacks realism and replicates the spatial layout of the training data, limiting the diversity.

On the other hand, large language models also play a crucial role. CuPL [31] leverages GPT-3 [32] to generate text descriptions that enhancing zero-shot image classification. CLOUDS [9] uses Llama [33] to create prompts for diffusion models.

However, they fail to adequately consider the complexities introduced by varying weather and lighting conditions. In contrast, our approach employs a chain of LLMs acting as agents to not only craft detailed descriptions of complex real-world scenarios but also implement a tailored generation strategy. This ensures that the generated images are both diverse and realistic, and address the class imbalance problem in challenging conditions.

III. Proposed Method

A. Overview

WeatherDG aims to generate images tailored to weather-specific autonomous driving scenes, enhancing semantic segmentation in challenging conditions. We begin by fine-tuning a diffusion model to adapt scene priors from the source domain, ensuring the generated images are within a driving scene (Section III-B). Next, we employ a procedural prompt generation method to create detailed prompts that enable the diffusion model to produce realistic and diverse weather and lighting effects (Section III-C). Additionally, we incorporate a probability-oriented sampling strategy for prompt generation. Following this, we use UDA training methods to leverage

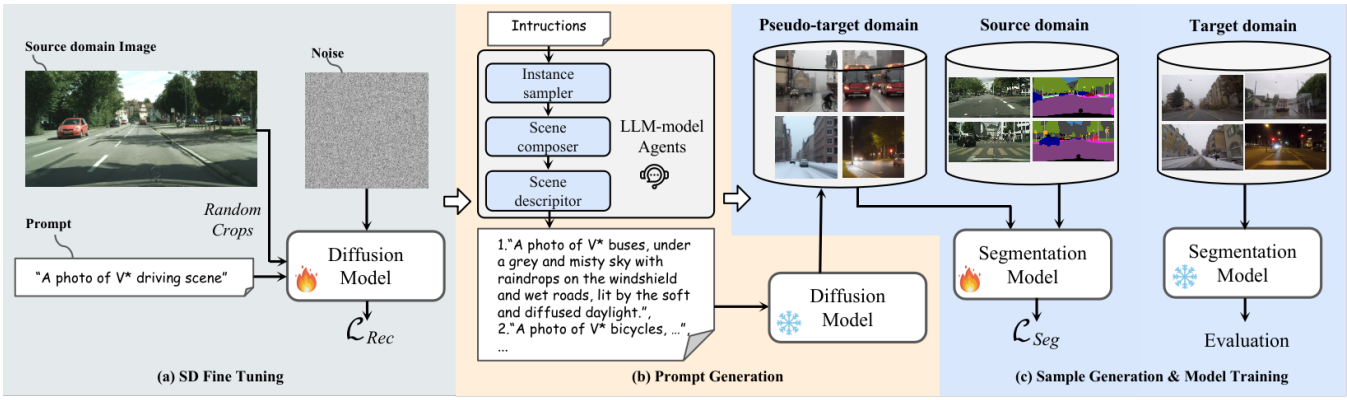


Fig. 3: **The overview of WeatherDG pipeline.** (a) We first fine-tune a text-to-image diffusion model to integrate scene priors from the source domain. This ensures the images generated by the diffusion model accurately depict driving scenes. (b) Next, we employ a chain of LLM agents to procedurally construct detailed prompts that can enrich tailed classes and generate diverse weather and lighting effects with the fine-tuned model. (c) After generating images with these prompts, we utilize UDA techniques to train these images with the source domain dataset, followed by evaluation on real-world target datasets.

these generated images for semantic segmentation training (Section III-D). The overview of the proposed approach is shown in Figure 3.

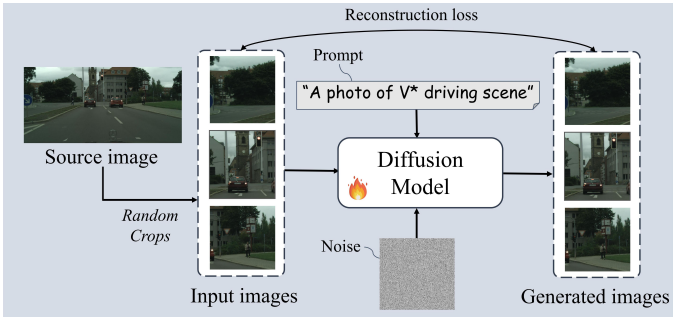
B. SD Fine-tuning

Recently, diffusion models have advanced the field of generative domain adaptation, showcasing exceptional capabilities in producing photo-realistic images conditioned on text [30]. However, directly applying diffusion models in autonomous driving setting presents challenges due to shifts in scene priors like style and layout. For example, the model often generates artistic images or adopt a bird's eye view perspective, which is different from the images in real-world autonomous driving datasets, as shown in Figure 4b. When these images are incorporated during training, they can disrupt the knowledge the model has acquired from the labeled source domain, thereby harming the semantic segmentation performance. To address these issues, we finetune a diffusion model [16] to generate diverse images that retain content and layouts relevant to source domain. The input consists of a clean image paired with a corresponding prompt text, while the output is an image tailored to the autonomous driving domain.

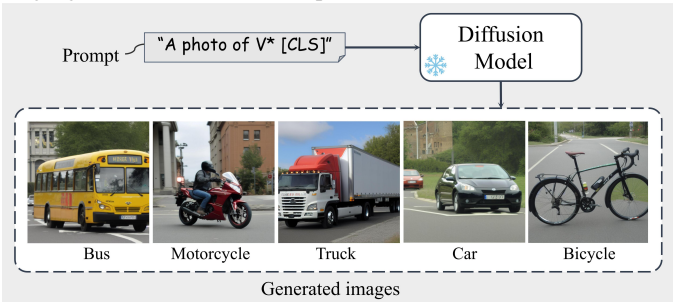
We follow similar text-to-image diffusion model training procedures described in the relevant work [27], [34]. Specifically, we use a unique identifier in the prompt to link priors to one single image choosen from the source domain dataset. For instance, the prompt "A photo of V^* driving scene" associates with image patches cropped from the selected image, where V^* identifies the scene, and "driving scene" describes it broadly. This prevents language drift and improves model performance [34]. The training procedure is presented in Figure 4a. After training, the model captures scene priors through the unique identifier V_* and can generate new instances in similar contexts within autonomous driving datasets, as shown in Figure 4b.

C. Procedural Prompt Generation

To enable the diffusion model to generate weather and lighting effects, it is essential to integrate specific weather



(a) Training stage: A unique identifier, V^*, links prompts to specific image patches from the dataset, following related work to maintain language-scene association and prevent drift.



(b) Inference stage: After training, the model uses the identifier V^* to generate new instances in autonomous driving contexts, ensuring coherent scene synthesis.

Fig. 4: The detailed process of Stable Diffusion [16] fine-tuning.

conditions and times of day into the prompt. However, simplistic templates such as "A photo of [CLS], [WEATHER], [TIME]" often do not yield diversity and details, necessitating the need for more nuanced descriptions. Manually writing these descriptions is labor-intensive; therefore, we consider adopting LLM models to automate this process. Additionally, given the scarcity of dynamic object samples in adverse weather conditions, we need to consider a balanced generation strategy during the prompt generation to enrich these objects. To this end, our requirements for the generated prompts are threefold: 1) it should incorporate a balanced generation strategy, 2) introduce different weather and lighting conditions in an even manner, and 3) provide detailed descriptions of

these conditions. Most importantly, all of the prompts are automatically generated by an LLM model to reduce human effort. During the implementation, we found that directly giving a single instruction for one LLM model fails to meet all of our requirements. Specifically, the generated prompts frequently do not align with our designed generation strategy or achieve the level of detailed description we expect. To address this, we develop a procedural prompt generation method involving a sequence of three LLM agents—namely, instance sampler, scene composer, and scene descriptor. This hierarchical approach enables us to generate precisely tailored text prompts for image generation, ensuring each aspect of the prompt aligns with our intended outcomes.

a) Instance sampler: During the inference phase, we can generate a specific instance by simply adding the instance name <CLS >to the prompt, which serves as the task for the instance sampler. However, we notice that the semantic segmentation evaluation metrics for "thing" classes are significantly lower compared to "stuff" classes in most autonomous driving datasets. This disparity can be attributed to the class imbalance problem where "stuff" classes have more occurances in the images than "thing" classes. In particular, this issue is exacerbated under adverse weather conditions , as dynamic objects are less frequently present on the road in snowy or nighttime scenes.

Therefore, we aim to enhance the instance sampler agent with the capability to employ a probability-oriented sampling strategy, giving underrepresented classes under adverse weathers higher sampling probabilities. This increases the likelihood of these rare classes presented in the generated images. In detail, we first compute the semantic label distribution of i-th thing classes E_i in a typical adverse weather dataset [35] by:

$$E_i = \frac{D_{\text{i}}}{D_{\text{thing}}}, \tag{1}$$

where D_{i} represents number of semantics labels of each thing class,D_{thing} stands for number of all thing classes. Then the sampling probability P_{i} can be formulated as:

$$P_i = \frac{1}{\sum_{j=1}^{n} E_j} \times \frac{1}{E_i}. \tag{2}$$

Then, we allow the instance sampler agent to query the sampling probability to generate prompts. In this way, we can alleviate the shortage of instances that rarely appear in adverse weather conditions.

b) Scene composer: Intuitively, to enable the generation of images covering a wide range of weather effects and different times of the day, we can design the prompt template as 'A photo of <CLS >, <WEATHER >, <TIME >' to describe the image to be generated. Here, <CLS > refers to the classes in the typical autonomous driving dataset, <WEATHER > specifies the weather conditions, and <TIME > indicates the time of the day. To enhance the balance and diversity of the generated dataset, we include three common weather conditions: snowy, rainy, and foggy, along with two distinct times of day: daytime and nighttime. Each condition and time period is equally represented in the dataset to ensure

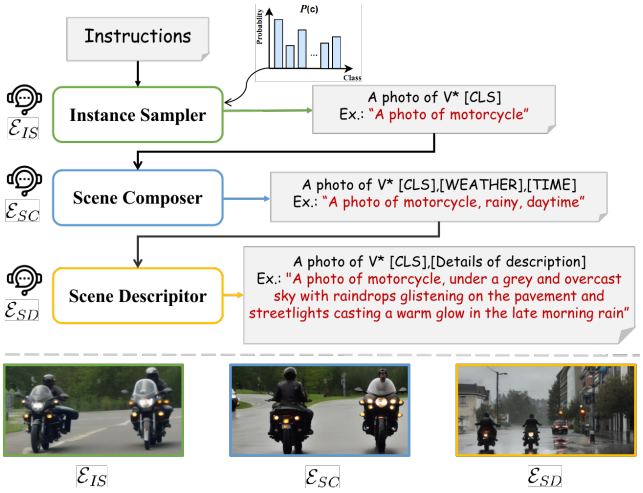


Fig. 5: The process of prompt generation by gradually introducing LLM agents: $\mathcal{E}_{IS}$, $\mathcal{E}_{SC}$ and $\mathcal{E}_{SD}$. The images correspond to the results generated using prompts created by different LLM agents.

comprehensive coverage and variability. However, merely incorporating categories of weather and time category does not guarantee the desired diversity in the effects, it often generates a singular subject with minimal environmental context and a limited range of effects in relatively simple scenes. As shown in Figure 5, with the prompt "A photo of motorcycle, rainy, daytime" generated by $\mathcal{E}_{SC}$, the model adds reflections on the road in daytime lighting, indicating a humid, rainy day, but the effect is subtle and the scene details are limited. Although the result does not fully meet expectations, it provides a solid foundation for further improvement.

c) Scene descripitor: To ensure the generated images reflect a variety of environmental effects and capture intricate details that resemble real-world scenes, we found that providing detailed descriptions of weather conditions and lighting is useful. For example, for the prompt "A photo of a motorcycle in rainy, daytime", we enhance the "rainy" aspect by including details such as "under a grey and overcast sky with raindrops on the pavement". Similarly, for "daytime", we add specifics like "streetlights casting a warm glow in the late morning rain". Using the crafted prompt, the model incorporates elements like buildings, streetlights, a reflective road, and a misty sky into the scene, significantly enhancing the complexity and realism of the generated images. This enables the model trained with these images to generalize better to real-world scenarios.

However, manually crafting these detailed descriptions can be labor-intensive. By using LLM, we can automatically generate these nuanced elements, making the creation of highly detailed scenes more efficient. Consequently, we incorporate another LLM agent that receives prompts produced by $\mathcal{E}_{SC}$ and uses the provided information to generate various detailed descriptions of scenes. In this way, we can enable a diversity of weather and lighting effects generation.

d) Procedural generation: The procedural prompt generation pipeline is as follows: first, the instance sampler $\mathcal{E}_{IS}$ queries the sampling probability to select the class to be

generated. Next, based on the instance sampler's output, the scene composer $\mathcal{E}_{SC}$ composes scenes in the prompt that include typical weather and times of the day. Finally, after receiving the prompt generated by $\mathcal{E}_{SC}$, the scene descriptor $\mathcal{E}_{SD}$ creates detailed descriptions of the corresponding weather and lighting conditions. Taking the probability of the instance sampling as $\mathcal{P}$, the process of generating text prompt $\mathcal{T}$ can be formulated as:

$$\mathcal{T} = \mathcal{E}_{SD}\left(\mathcal{E}_{SC}\left(\mathcal{E}_{IS}(\mathcal{P})\right)\right).$$

D. Sample Generation & Model training

With the prompt $\mathcal{T}$, we send it to the fine-tuned diffusion model to generate image samples. Although the model can produce realistic images featuring challenging weather and lighting conditions, it is still difficult to incorporate pixel-level information because the images lack semantic labels. To overcome this, we employ UDA methods such as DAFormer [1] and HRDA [36], which facilitate the adaptation of a segmentation model to an unlabeled target dataset. Specifically, we train these methods using Cityscapes [37] dataset as the source domain, and our generated dataset as the pseudo-target domain. The model is then evaluated on the target three real-world benchmarks [35], [38], [39]. The proposed framework can be adapted to UDA method easily, transforming it into a domain generalization method.

IV. Experiments

A. Dataset

We conduct the experiments with domain generalization settings, using Cityscapes [37] as source domain dataset, ACDC [35], BDD100k [38] and DarkZurich [39] as test target domain dataset. The Cityscapes dataset comprises 2,975 images captured under standard weather conditions and during daytime, each accompanied by corresponding semantic segmentation labels for training. The ACDC dataset features images taken under various typical weather and lighting conditions, including snowy, rainy, foggy, and nighttime settings. BDD100k contains various weathers and geographic locations. DarkZurich consists of images captured at night, providing a challenging environment for nighttime visual perception tasks. For fine-tuning the diffusion model, we use a single image sampled from the Cityscapes dataset. For unsupervised domain adaptation (UDA) training, we utilize the Cityscapes training set along with images generated by our diffusion model. For evaluation, we use 406 images from the ACDC validation set, 1,000 images from the BDD100k validation set, and 50 images from the DarkZurich validation set.

B. Implementation details

We utilize the Stable Diffusion [16] as pretrained model and fine-tune it using DreamBooth [34] method. During fine-tuning stage, we randomly crop a patch of 512×512 size from the images selected from Cityscapes dataset. Then we pair it with customized prompt "a photo of V driving scene". We use mean square error loss to quantify the differences between the original image patch and the generated image during model training. After training, we use text prompts generated by a chain of Llama [33] models to create images. The generated images are further used in UDA training for generalizing normal weather source domain to adverse conditions domain.

V. Evaluation and Discussion

A. Comparison with State-of-the-art

We evaluate WeatherDG against state-of-the-art domain generalization models using ResNet-50 [40] and MiT-B5 [41] as encoders. As shown in Table I, for models using ResNet-50 as encoder, our model consistently outperforms state-of-the-art methods with the highest average mIoU score. With MiT-B5 as backbone, our model exceeds the second-best model MIC [10] on the ACDC and DarkZurich datasets by over 10% and on the BDD100K dataset by 4.3% in mIoU performance, achieving the best semantic segmentation performance. In addition, we visualize our model's semantic segmentation results under challenging conditions and compare them with MIC. As shown in Figure 1, our model correctly segments sidewalks and sky in foggy and nighttime scenes. In the rainy scene, reflections on the road that are falsely recognized as vehicles by MIC are alleviated by our model. In the snowy scenario, our model successfully detects pedestrians on the road that MIC fails to recognize. These findings indicate our model's superior generalization capabilities compared to state-of-the-art methods in real-world challenging weather and lighting conditions.

Method	Encoder	Test domains mIoU			Avg.
		ACDC	BDD100K	DarkZurich	
Source-only	ResNet-50	35.9	37.3	9.0	27.4
IBN-Net [42]		42.0	45.8	17.3	35.0
RobustNet [7]		41.7	43.4	19.4	34.8
SHADE [20]		42.1	**49.1**	22.6	37.9
DPCL [43]		43.8	44.9	23.4	37.3
WeatherDG (ours)		**45.2**	45.8	**23.5**	**38.2**
Source-only	MiT-B5	44.6	44.7	18.7	36.0
DAFormer [1]		47.8	49.2	24.7	40.6
HRDA [36]		46.3	49.8	25.8	40.6
MIC [10]		50.0	53.1	21.5	41.5
WeatherDG (ours)		**60.2**	**57.4**	**35.3**	**51.0**

TABLE I: **Domain generalization performance (mIoU (%)) of state-of-arts methods using ResNet-50 and MiT-B5 as encoders**. The compared methods are retrained with Cityscapes dataset. Evaluations are performed on ACDC, BDD100K, and DarkZurich datasets that feature adverse conditions, such as snow, rain, fog, and low-light scenarios.

B. Influence of UDA methods for training

Method	Encoder	Test domains mIoU			Avg.
		ACDC	BDD100K	DarkZurich	
DAFormer [1]	ResNet-50	**45.2**	**45.8**	**23.5**	**38.2**
MIC [10]		43.8	43.7	22.6	36.7
HRDA [36]		44.5	44.6	22.4	37.2
DAFormer [1]	MiT-B5	53.3	53.5	24.7	43.8
MIC [10]		60.0	54.4	32.6	49.0
HRDA [36]		**60.2**	**57.4**	**35.3**	**51.0**

TABLE II: **Comparison of mIoU performance of UDA methods** trained using the labeled Cityscapes dataset as the source dataset and our generated unlabeled images as the target dataset.

Fig. 6: **Comparison of images generated by the plain stable diffusion model (top row) and our fine-tuned model (bottom row) using the same prompt template**. The results illustrate that our fine-tuned model significantly reduces artistic and unrealistic elements, generating images more aligned with real-world autonomous driving scenarios.

We investigate the influence of UDA methods by utilizing three different state-of-the-art approaches including DAFormer [1], HRDA [36], and MIC [10], each trained with Cityscapes and our generated dataset. As shown in Table II, DAFormer demonstrates superior adaptation to the pseudo-target domain among the ResNet-50 models, while HRDA achieves the best generalization performance across the three domains with the MiT-B5 encoder. To further study the influence of these methods under different conditions, we evaluate them across four typical challenging scenarios in the ACDC dataset. As indicated in III, all methods perform best in foggy conditions and worst in nighttime conditions. Notably, none of these methods exceed a 40% performance in nighttime scenes, which is significantly lower than in other scenarios. This could be due to the substantial appearance differences between nighttime scenes and the predominantly daytime images in the training source domain, making adaptation challenging. This finding suggests that simply adding a pseudo-target dataset for adaptive training may be inadequate for complete knowledge transfer in the nighttime domain, necessitating more advanced adaptation techniques.

Method	Test weathers				Avg.
	Snow	Foggy	Rainy	Nighttime	
Source-only	49.9	61.9	47.8	19.6	44.8
DAFormer [1]	54.2	66.8	54.1	27.3	50.6
MIC [10]	59.4	74.2	62.0	36.9	58.0
HRDA [36]	**59.7**	**75.9**	**64.6**	**38.7**	**59.7**

TABLE III: Comparison of mIoU performance of UDA techniques across typical weather and lighting conditions.

C. Influence of SD Fine-tuning.

To demonstrate the influence of scene prior adaptation, we compare images generated by the plain stable diffusion model and our fine-tuned model. We use the same prompt, templated as "A photo of [CLS]" for each model to generate commonly seen objects in the autonomous driving dataset. The results in Figure 6 show that the plain stable diffusion tends to generate images with an artistic style or cinematic photography effects, as seen with "truck," "bicycle," "motorcycle," and "bus". For "car" and "train", the images present different camera perspectives, such as bird's-eye view. Additionally, for "traffic light", "traffic sign", and "person", the model exhibits excessive creativity, generating overly stylized traffic lights and rendering "person" as sketches. For the "rider" category, the stable diffusion model imagines a surfer on the sea.

These samples may acquire incorrect pseudo-labels that negatively impact adaptive training and harm segmentation performance. In contrast, our fine-tuned model ensures that the generated instances are contextually appropriate for the driving scene, resulting in more accurate pseudo-labels and better adaptation to the autonomous driving domain. As shown in Table IV, the performance of semantic segmentation model trained with images generated by the fine-tuned model is better than plain stable diffusion.

Method	ACDC	BDD100K	DarkZurich
w/o Fine-tuning	49.9	52.1	24.7
w Fine-tuning	**50.8**	**53.2**	**25.6**

TABLE IV: Comparison of mIoU performance of models trained on datasets generated with and without fine-tuned model.

D. Influence of Procedural Prompt Generation

To illustrate the effectiveness of our procedural prompt generation, we compare images generated by text prompts created by different LLM agents in Figure 7. The results show that the instance sampler can generate the desired objects with a basic prompt such as "A photo of [CLS]". By specifying the general category of weather and time of day, the scene composer only adds basic weather effects to the image, though these effects are subtle. Additionally, the generated samples primarily focus on the subject, often lacking scene details. When the scene descriptor crafts more detailed scene descriptions in the prompt, the model produces images with various realistic weather and lighting effects. As shown in third row, for snowy weather, the model generates a complex scene with heavy accumulated snow on the road and even snowflakes in the air. For rainy weather, the prompt generates reflections, raindrops, and a misty effect, indicating heavy rain. For nighttime, we can see that $\mathcal{E}_{SC}$ fails to add sufficient nighttime lighting, but with prompts generated by $\mathcal{E}_{SD}$, the scene in the image exhibit a darker tone and more intricate details, creating a more realistic nighttime environment. In addition, we evaluate the mIoU performance of the DAFormer [1] trained with datasets generated using different LLM agents. In Table V, the results demonstrate that progressively refining

(a) Snowy (b) Foggy (c) Rainy (d) Nighttime

Fig. 7: **Comparison of images generated using prompts created by different LLM agents.** Each row represents images generated by a specific LLM agent, while each column showcases images generated by different LLM agents for specific weather or lighting effects during the procedural prompt generation. The results demonstrate that $\mathcal{E}_{IS}$ (top row) enables the model to generate diverse instances, albeit with limited scene detail. While $\mathcal{E}_{SC}$ (middle row) allows the model to generate weather and lighting effects, the overall impact is rather subtle. With detailed descriptions crafted by $\mathcal{E}_{SD}$, the model (bottom row) produces intricate scene details and more diverse weather and lighting effects, significantly enhancing the variety and realism of the generated images.

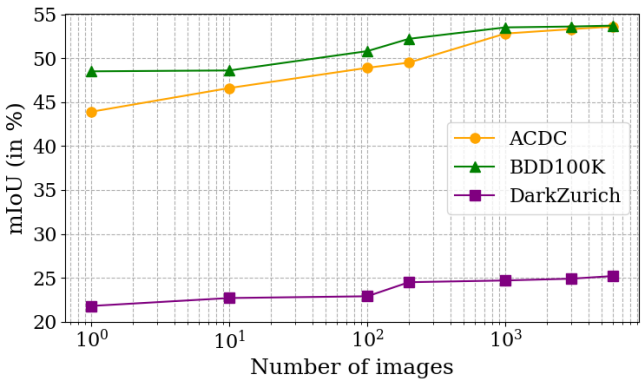


Fig. 8: Impact of the number of generated images on training model's mIoU performance.

the base prompt "A photo of [CLS]" using these LLM agents results in higher mIoU scores.

Models	LLM-Agents			mIoU
	$\mathcal{E}_{IS}$	$\mathcal{E}_{SC}$	$\mathcal{E}_{SD}$	
M_{base}	–	–	–	50.8
M_1	✓	–	–	51.5
M_2	✓	✓	–	52.1
M_3	✓	✓	✓	53.3

TABLE V: Comparison of mIoU performance for models trained on datasets generated by introducing different LLM agents.

E. Influence of Numbers of Generated Images

To investigate the impact of the generated dataset size on the performance of the segmentation model, we evaluate the mIoU performance of the DAFormer model trained with the generated dataset on the ACDC, BDD100K, and DarkZurich datasets. As shown in Figure 8, a substantial performance gain are observed as the number of images increases from 10 to 1000 for all three datasets, after which the performance gains level off. Notably, the Dark Zurich dataset shows much smaller improvements with the increased number of images compared to the other two datasets. This can be attributed to the inherent learning difficulty for nighttime images during the model training. These findings indicate that: 1) while increasing the amount of training data generally enhances model performance, the benefits may plateau after some point; and 2) inherent dataset characteristics can impact the extent of the model's improvement.

VI. Conclusion

In this paper, we present WeatherDG, a novel approach for domain generalization in semantic segmentation under adverse weather conditions. By combining Stable Diffusion (SD) with a Large Language Model (LLM), our method enables automated generation of realistic images resembling real-world driving scenarios. Fine-tuning SD, along with procedural prompt generation and a balanced strategy, creates diverse weather effects and enhances tailed classes in generated images. These images, combined with source data, improve model generalization. Experiments across challenging datasets show that WeatherDG significantly boosts semantic segmentation performance, setting a new benchmark for robustness in autonomous driving.

References

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

[2] A. Seppänen, R. Ojala, and K. Tammi, "4denoisenet: Adverse weather denoising from adjacent point clouds," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 456–463, 2023.

[3] Y. Lee, Y. Ko, Y. Kim, and M. Jeon, "Perception-friendly video enhancement for autonomous driving under adverse weather conditions," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7760–7767.

[4] Y. Lee, J. Jeon, Y. Ko, B.-G. Jeon, and M. Jeon, "Task-driven deep image enhancement network for autonomous driving in bad weather," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 746–13 753, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238857297

[5] X. Jiang, J. Huang, S. Jin, and S. Lu, "Domain generalization via balancing training difficulty and model capability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 993–19 003.

[6] D. Peng, Y. Lei, M. Hayat, Y. Guo, and W. Li, "Semantic-aware domain generalized segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2594–2605.

[7] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 580–11 590.

[8] J. Niemeijer, M. Schwonberg, J.-A. Termöhlen, N. M. Schmidt, and T. Fingscheidt, "Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 2830–2840.

[9] Y. Benigmim, S. Roy, S. Essid, V. Kalogeiton, and S. Lathuilière, "Collaborating foundation models for domain generalized semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 3108–3119.

[10] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[12] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3707478

[13] R. Zurbrügg, H. Blum, C. Cadena, R. Siegwart, and L. Schmid, "Embodied active domain adaptation for semantic segmentation via informative path planning," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8691–8698, 2022.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[15] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[17] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "Robust place categorization with deep domain generalization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2093–2100, 2018.

[18] J. Weyler, T. Läbe, F. Magistri, J. Behley, and C. Stachniss, "Towards domain generalization in crop and weed segmentation for precision farming robots," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3310–3317, 2023.

[19] J. Niemeijer and J. P. Schäfer, "Domain adaptation and generalization: A low-complexity approach," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 1081–1091. [Online]. Available: https://proceedings.mlr.press/v205/niemeijer23a.html

[20] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. H. Lee, "Style-hallucinated dual consistency learning for domain generalized semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[21] T. Sun, C. Lu, T. Zhang, and H. Ling, "Safe self-refinement for transformer-based domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7191–7200.

[22] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4489–4495.

[23] M. Diaz-Zapata, Ö. Erkent, and C. Laugier, "Instance segmentation with unsupervised adaptation to different domains for autonomous vehicles," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2020, pp. 421–427.

[24] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[25] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 433–443.

[27] Y. Benigmim, S. Roy, S. Essid, V. Kalogeiton, and S. Lathuilière, "One-shot unsupervised domain adaptation with personalized diffusion models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 698–708.

[28] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. de Charette, "Pøda: Prompt-driven zero-shot domain adaptation," in *ICCV*, 2023.

[29] A. Kerim, F. Chamone, W. L. Ramos, L. S. Marcolino, E. R. Nascimento, and R. Jiang, "Semantic segmentation under adverse conditions: A weather and nighttime-aware synthetic data-based approach," in *33nd British Machine Vision Conference 2022, BMVC 2022*, 2022.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[31] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 691–15 701.

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[34] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[35] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 765–10 775.

[36] L. Hoyer, D. Dai, and L. V. Gool, "Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 01, pp. 220–235, jan 2024.

[37] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[38] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[39] 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*, 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[42] J. S. Xingang Pan, Ping Luo and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018.

[43] L. Yang, X. Gu, and J. Sun, "Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 10 789–10 797, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26280