CaRS: A Causal Intervention Segmentation Framework and Benchmark Dataset for Autonomous Driving under Transitional Weather Conditions

Supplementary Material

In the supplementary material, we present additional information not included in the main paper due to space limitations. We have meticulously organized the details into individual sections to enhance clarity and facilitate a comprehensive understanding of our work. Table 1 provides a summary of adopted notations.

1. Related Work

This section discusses recent works on semantic & instance segmentation for AVT in adverse weather conditions and causal learning-based scene perception models.

1.1. Semantic Segmentation in Adverse Weather Conditions

Pixel-wise labeling for AVT in adverse weather conditions presents a significant challenge due to low visibility [24, 26]. Lee *et al.* [17] presented FIFO to learn fog-invariant features for foggy scene segmentation using a fogpass filtering module. Shaik *et al.* [28] presented IDD-AW, a large-scale dataset for semantic segmentation in adverse weather conditions. Franchi *et al.* [12] introduces the InfraParis multimodal dataset that supports multiple tasks across three modalities: RGB, depth, and infrared. Li *et al.* [18] proposed VBLC, a framework for normal-to-adverse adaptation, aimed at eliminating the need for reference images. Zhou *et al.* [42] introduced attention-based cross-view transformers designed for map-view semantic segmentation across multiple cameras. Liao *et al.* [20] introduced a pseudo-label diffusion method for unsupervised domain adaptation for semantic segmentation. Zhou *et al.* [43] introduced a self-adversarial disentanglement framework to reduce the domain gap caused by weather conditions for semantic segmentation. Cai *et al.* [6] proposed a semantic segmentation model based on the multi-target pan-class intrinsic relevance. Wang *et al.* [33] proposed a backpropagation-free approach for domain adaptive semantic segmentation, called Dynamically Instance-Guided Adaptation (DIGA). Their method combines a Distribution Adaptation Module (DAM) and a Semantic Adaptation Module (SAM) to jointly optimize the model across feature distributions and semantic representations. On the other hand, Zheng *et al.* [41] developed a deep inference network for road area and road element segmentation. Bruggemann *et al.* [5] present CMA, a model adaptation method for cross-condition semantic segmentation. CMA leverages image-level correspondences to learn condition-invariant features through a contrastive loss. Kothandaraman *et al.* [16] proposed a source-free domain adaptation for road segmentation in adverse weather conditions.

While the above approaches are notable, they largely operate within discrete weather categories and depend on domain adaptation for a limited set of predefined conditions. In contrast, our method adopts a causal intervention strategy that mitigates spurious correlations arising from transitional weather, enabling more robust segmentation.

1.2. Instance Segmentation in Adverse Weather Conditions

Unlike semantic segmentation, which labels each pixel with a class, instance segmentation assigns a unique label to each instance of an object in an image. Instance segmentation is advantageous in autonomous driving, where precise object representation is paramount [2]. Several works proposed instance segmentation models for AVT in inclined weather conditions [15, 33]. Yin *et al.* [35] devised Sem2Ins, a real-time instance segmentation model that generates instance boundaries from semantic segmentation labels using conditional GANs. Musat *et al.* [23] employed GAN and CycleGAN to create diverse weather conditions and conducted instance segmentation on the resulting dataset. Wang *et al.* [31] proposed CDAC, which leverages the consistency between self-attention and cross-domain attention predictions to address domain shifts at both the attention and output levels. Cheng *et al.* [10] introduced the SparseInst model that uses sparse instance activation maps to highlight essential regions for each object. Cheng *et al.* [9] introduced mask2former, a universal segmentation model that utilizes masked attention for all segmentation tasks, including instance segmentation. Table. 1 of the main paper presents the current literature on semantic and instance segmentation for AVT in adverse weather conditions. An observation from the table reveals a lack of emphasis on segmenting road elements in transitional weather conditions in existing works. The proposed work differs from previous methods in tackling semantic and instance segmentation simultaneously across transitional weather conditions. This approach empowers us to tackle the uncertainties brought by unpredictable weather patterns, offering a more resilient solution, particularly for segmentation in real-world settings.

1.3. Causal Scene Perception

Deep learning models are powerful but often lack transparency. Several works have been proposed to enhance understanding by integrating complementary strengths from various domains. Utilizing causal intervention has been central to this effort [32, 37]. Pourkeshavarz *et al.* [25] proposed a causal disentanglement approach for trajectory prediction that enhances model robustness and generalization by effectively separating causal from spurious factors. Venkataramani *et al.* [30] proposed Causal Feature Alignment (CFA), which mitigates spurious background features by leveraging spatial localization of causal features on a subset of training data, instead of relying on group labels. Zhang *et al.* [38] introduced a causal interventional reasoning module to obtain weather-invariant features in object detection. Wang *et al.* [32] presented CaaM, a causal attention module integrated into CNNs and transformers to improve classification performance by tackling confounding factors. Xu *et al.* [34] devised a domain generalization model with a multi-view adversarial discriminator and a spurious correlation generator to enhance object detection. While these scene perception models leveraging causality have demonstrated strong performance, there is a noticeable gap in causal intervention for segmentation tasks. To achieve this, Miao *et al.* [22] proposed a causal semi-supervised learning approach for medical image segmentation. Li *et al.* [19] introduced a causal interventional segmentation approach using transformers. However, these methods are unable to address confounding effects in autonomous driving, particularly in addressing confounding scenarios like transitional weather conditions. Developing models to tackle such complexities is essential for improving segmentation accuracy and robustness in practical autonomous driving applications. Hence, this paper introduces a novel causal road and rest segmentation method that significantly reduces the effects of transitional weather conditions on segmentation performance, thereby addressing a crucial challenge in AVT.

2. Preliminaries

2.1. Latent Space Representation

The latent space generated by the variational autoencoder (VAE), a hidden representation learned from input images, is shown in Figure 1. Unlike a traditional autoencoder, which creates discrete clusters by mapping input sequences directly to latent representations, the VAE produces a smooth and continuous latent space by learning a Gaussian distribution from the input data. This continuity and smoothness are key features of VAEs, enabling seamless interpolation between points in the latent space.

2.2. Data Interpolation

Interpolation in the VAE's latent space begins with selecting two points (z_i and z_j) corresponding to different input images. A series of intermediate latent vectors is then generated using linear interpolation. The VAE decoder transforms these intermediate vectors into a progressive sequence of images. In the context of weather images, this technique creates seamless transitions between distinct weather conditions, producing a continuous evolution of visual features. Figure 2 illustrates the latent data interpolation technique, which generates the new latent variable z by interpolating between the latent variables z_i and z_j. This interpolation is given by $z = \alpha z_i + (1 - \alpha) z_j$, for some $\alpha \in [0, 1]$.

Table 1. Summary of adopted notations

Notation	**Definition**
$\mathcal{X}$	Object feature representation
$\mathcal{Y}$	Output labels
$\mathcal{A}$	Training set
f	Backbone network
$\mathcal{M}$	Number of objects
d	Feature dimension
$\mathcal{R}$	Random number
$\mathcal{L}$	Label categories
$\mathcal{S}$	Spurious correlations
$do(.)$	Causal intervention
t	Transition sequence
T	Length of transition sequence
β_0 & β_1	Coefficients
e	Error
v_i & w_i	Input image pair
N	Size of training set
θ	Variational parameters
γ	optimization parameters
z	Latent variable
ψ	Model parameters
μ	Mean of gaussian distribution
σ	Variance of gaussian distribution
α	coefficient
$\mathcal{N}$	Normal distribution
$\hat{x}$	Interpolated image
$\mathcal{I}$	Covariance
$\mathcal{C}$	Causal features
a	Intermediate features
h	Residual block
$\widetilde{x}$	Fused features
$\mathcal{L}$	Loss
λ	weight factor
$\mathcal{D}$	Data Distribution

2.3. Transitional Weather Conditions

In continuation of Section 2 of the main paper, we discuss the importance and challenges of transitional weather for autonomous vehicle technology.

Importance of addressing weather transitions in autonomous vehicles. Transitional weather conditions significantly impact scene perception performance in autonomous driving. Although existing datasets such as Foggy Cityscapes and MultiWeatherCity capture adverse

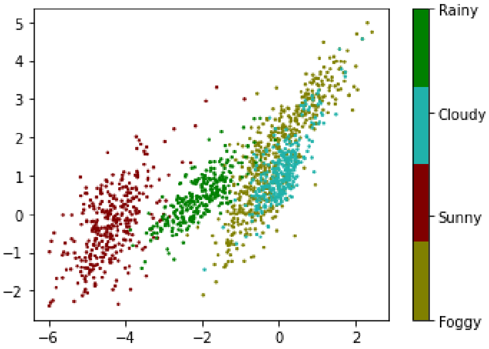


Figure 1. Latent space representation of the input images generated by VAE. The representation is smooth and continuous.

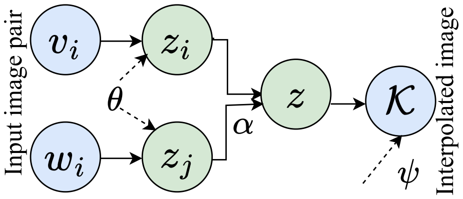


Figure 2. Latent data interpolation involves generating z_i and z_j from input pairs v_i and w_i using a VAE encoder. By interpolating z_i and z_j, we produce a new latent representation z, using which a new interpolated image is generated.

weather scenarios, they mainly focus on discrete conditions such as clear weather, heavy fog, or intense rain. However, real-world driving environments often involve gradual weather transitions (for example, cloudy to rainy or snowy to foggy). These transitions introduce additional complexities that go beyond those encountered in steady state weather conditions.

Unique challenges posed by transitional weather. Current datasets and models fail to address critical challenges associated with weather transitions: *(i)* Unlike static weather states, transitions involve dynamic changes in visibility and illumination, which can disrupt perception models trained on fixed conditions. *(ii)* Autonomous systems rely on stable environmental features for perception. However, weather transitions introduce inconsistencies in visual cues, leading to misclassification or unreliable predictions.

Limitations of existing models in handling transitional weather. Most perception models are trained on datasets with fixed weather conditions, making them ill-equipped to handle gradual transitions. Their primary limitations include: *(i)* Models trained on clear or extreme weather conditions lack the adaptability to generalize across intermediate weather changes, *(ii)* existing models often rely on static images or short sequences, missing the gradual evolution of visual features over time, *(iii)* during transitions, environmental features may not match any single weather condition category, causing prediction errors. These limitations reduce model robustness in real-world deployments, where smooth weather transitions are common. Addressing these gaps is essential for developing more reliable and adaptable autonomous vehicle perception systems.

Unique advantages of using VAE data interpolation for transitional weather generation. VAE interpolation for transitional weather generation offers several unique advantages. Leveraging the structured and continuous latent space of VAEs enables smooth and controllable transitions between distinct weather states (e.g., cloudy to rainy) while preserving the semantic content of the scene. This interpolation supports gradual appearance changes without altering critical features such as road layout and object boundaries, making it particularly effective for training robust segmentation models in the presence of spurious weather correlations. Additionally, VAEs can operate efficiently with limited data and inherently model uncertainty in the generation process, making them well-suited for capturing diverse and stochastic weather transitions. Compared to GAN- or diffusion-based methods, VAE interpolation produces more stable outputs with interpretable latent controls, thereby facilitating data generation for robust segmentation under transitional weather conditions.

2.4. Causal Intervention

Let $\mathcal{X} \in \mathcal{R}^{\mathcal{M}_d}$ be the object feature representations learned by a backbone network $f(.)$ from training images $\mathcal{A}$ where $\mathcal{X} = f(\mathcal{A})$ and $\mathcal{M}$ is the number of objects of dimension d. The corresponding labels are represented by $\mathcal{Y} \in \mathcal{R}^{\mathcal{L}}$, where $\mathcal{L}$ is the number of label categories. Traditional Bayesian models primarily focus on computing the likelihood $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ to predict object categories. However, these models often face challenges due to spurious correlations S between object features and their categories, impacting their predictive performance. Consequently, they struggle to capture the genuine causality necessary for accurately predicting certain categories. The formulation of likelihood can be expressed as:

$$\mathcal{P}(\mathcal{Y}/\mathcal{X}) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y}/\mathcal{X}, \mathcal{S})\mathcal{P}(\mathcal{S}|\mathcal{X}), \tag{1}$$

where $\mathcal{P}(\mathcal{S}|\mathcal{X})$ indicates the observational bias. To mitigate feature bias introduced by spurious correlations through $\mathcal{S} \to \mathcal{Y}$, it is crucial to utilize causal intervention, denoted as $\mathcal{P}(\mathcal{Y}|do(\mathcal{X}))$, rather than relying solely on the likelihood $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ when learning feature representations.

$$\mathcal{P}(\mathcal{Y}/do(\mathcal{X})) = \sum_{\mathcal{S}} \mathcal{P}(\mathcal{Y}/\mathcal{X}, \mathcal{S})\mathcal{P}(\mathcal{S}). \tag{2}$$

Here, $\mathcal{P}(\mathcal{S})$ denotes the prior probability of the confounding factors. Consequently, the prediction of $\mathcal{Y}$ is designed

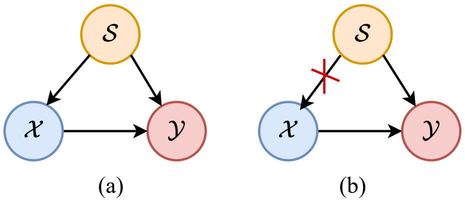


Figure 3. Visual representation of SCM: (a) Existing method: Confounder $\mathcal{S}$ influences the outcome. (b) Proposed method: Spurious correlations are eliminated between $\mathcal{S}$ and $\mathcal{X}$.

to include the confounding prior $\mathcal{P}(\mathcal{S})$ for the establishment of a robust segmentation model. In contrast to Eq. 1, the confounding factors $\mathcal{S}$ exhibit no correlation with the training samples $\mathcal{X}$. For example, $\mathcal{S}$ can be interpreted as the transitional weather conditions, e.g., cloudy to rainy, with a transition from distinct cloudy to rainy conditions. The category classifier $\mathcal{P}(\mathcal{Y}|\mathcal{X}, \mathcal{S})$ captures discriminative object features across different transitional weather conditions. This approach aims to mitigate the confounding bias arising from varying intensities of weather conditions in the training samples.

2.5. Structural Causal Model (SCM)

We introduce a causal formulation for the segmentation task by utilizing SCM to understand how confounding factors influence segmentation models, as shown in Figure 3. In SCM, we establish causal relationships among observed objects ($\mathcal{X}$), a set of confounding factors such as transitional weather conditions ($\mathcal{S}$), and observed labels ($\mathcal{Y}$). The causal connections between two nodes are denoted by $\rightarrow$. In Figure 3a, the notation $\mathcal{S} \rightarrow \mathcal{X}$ signifies that context $\mathcal{S}$ influences the performance of the segmentation model. Specifically, $\mathcal{S}$ produces both noisy features $\mathcal{X}_s$ and core object features $\mathcal{X}_o$, collectively forming a feature representation of $\mathcal{X}$. The noisy features refer to object-irrelevant attributes introduced by weather conditions, reducing the efficiency of segmentation models $\mathcal{X} \rightarrow \mathcal{Y}$. The object factors within $\mathcal{X}$ directly influence $\mathcal{Y}$, suggesting that core object attributes such as textures and shapes play a determining role in defining the object label. This causation remains unaffected by external factors like weather conditions, camera viewpoints, or backgrounds, as denoted by $\mathcal{X} \leftarrow \mathcal{S} \rightarrow \mathcal{Y}$. The notation $\mathcal{S} \rightarrow \mathcal{Y}$ highlights the influence of the confounding factors $\mathcal{S}$ on $\mathcal{Y}$.

We aim to establish the true causality between $\mathcal{X}$ and $\mathcal{Y}$ to develop an invariant feature representation capable of reducing the effect of spurious correlations. However, existing segmentation models [9, 10] focus on computing the likelihood $\mathcal{P}(\mathcal{Y}|\mathcal{X})$, fail to capture true causal relationships as shown in Figure 3a. To mitigate sampling bias from spurious correlations via $\mathcal{S} \rightarrow \mathcal{X}$, we use causal intervention $\mathcal{P}(\mathcal{Y}|do(\mathcal{X}))$ instead of likelihood $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ as shown in Figure 3b. Here, the $do(.)$ operator is used to identify causal effects from the given SCM.

3. Implementation Details

3.1. NWGM Implementation

For a given input $\mathcal{X}_i$ under multiple weather transitions $\mathcal{S}_t$, where $t = 0$ to T, we compute the model logits $f(\mathcal{X}_i, \mathcal{S}_t)$ for each t, then average as, $\bar{f}(\mathcal{X}_i) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} f(\mathcal{X}_i, s_t)$. The final prediction is obtained as, $\text{Softmax}(\bar{f}(\mathcal{X}_i))$. As Section 3.2 (**main paper**) mentions, this approximation allows us to move the expectation inside the softmax, reducing bias and variance.

3.2. Weather Invariant Loss

Figure 3 illustrates the causal intervention framework based on backdoor adjustment, with the corresponding intervention formula provided in Eq. 2. A standard cross-entropy loss alone is insufficient to realize this intervention. The second term in Eq.5 of the main paper is essential as a regularization component during training to enforce a shared, weather-invariant representation for segmentation across varying weather intensities. This regularization term is omitted at inference time, and only the intervened prediction is used. The weather invariant loss is designed to learn a representation, specifically the causal feature that remains consistent and robust for prediction across various intensities of weather conditions of the entire sequence T. It is given as

Definition 3.1 ***Weather invariant representation.*** *A representation $\mathcal{CA}(x)$ is said to be invariant across a weather sequence T if there exists a classifier f such that for all $t \in T$, $f \in \arg\min_{\bar{f}} \mathbb{E}[\text{mCE}(\bar{f}, \mathcal{CA}(x), t)]$, where* mCE *denotes the cross-entropy loss.*

This condition is enforced through the following optimization objective:

$$\begin{aligned} \min_{\mathcal{CA}, f} & \sum_{t \in T} \text{mCE}(f, \mathcal{CA}(x), t) \\ \text{subject to} \quad & f \in \arg\min_{\bar{f}} \text{mCE}(\bar{f}, \mathcal{CA}(x), t), \ \forall t \in T. \end{aligned} \tag{3}$$

The definition of mCE is provided in Section 3.2 of the main paper. Intuitively, the above objective performs empirical risk minimization on $\mathcal{CA}(x)$, while enforcing that the learned representation remains invariant across T. However, this formulation leads to a bi-level optimization problem, where each constraint involves solving an inner-loop minimization, making it computationally expensive and challenging in practice. To address this, [1] proposes a practical relaxation of the objective, expressed as:

$$\min_{\mathcal{CA},f} \sum_{t\in T} \mathrm{mCE}(f,\mathcal{CA}(x),t) + \lambda \left\|\nabla_f \mathrm{mCE}(f,\mathcal{CA}(x),t)\right\|_2^2 . \tag{4}$$

In our work, we adopt this formulation, Eq. 4, as the weather invariant loss, and incorporate it in Eq.6 and Eq.7 of the main paper. Algorithm 1 summarises the overall training pipeline.

Algorithm 1 Adversarial CaRS Training

```
Require: Dataset $\mathcal{A}$
Ensure: Representations $\mathcal{CA}$, $\mathcal{SA}$
1: Initialize model $f$ and representations $CA$, $SA$
2: for $i = 1$ to $n$ do
3:     for each $x \in \mathcal{A}$ do
4:         $c_i \leftarrow \mathcal{CA}_i(x)$                // Causal feature
5:         $s_i \leftarrow \mathcal{SA}_i(x)$     // Spurious/confounding feature
6:     end for
7:     Update $f_i, \mathcal{CA}_i, \mathcal{SA}_i$ using Eq. 6 of the main paper
       with $t$                                  // Min-game
8:     Update $\gamma, t$ using Eq.7 of the main paper          //
       Max-game
9: end for
```

4. Software and Machine Setup

We used Python 3.8, PyTorch, and TensorFlow for our experiments on an NVIDIA Tesla M60 8GB GPU. To perform semantic segmentation, our CaRS model, trained for 50 epochs with a batch size of 16 and a learning rate of 0.001, minimizes the binary cross-entropy loss function using the Adam optimizer. In addition, to perform instance segmentation, our CaRS model, trained for 30 epochs with a batch size of 4 and a learning rate of 0.005, minimizes the cross-entropy loss function using the SGD optimizer with momentum. Also, our SCG model trained for 100 epochs with a batch size of 16 and a learning rate of 0.0001, minimizes the KL divergence loss function using the Adam optimizer. We randomly split the dataset into 70% training, 10% validation, and 20% testing sets. The baseline semantic and instance segmentation models follow the same training setup as CaRS. The hyperparameter λ in the weather invariant loss was set to either 5×10^4, following the settings used in [29]. λ was fixed at 1×10^6 for the max-game. The number of iterations n was varied between 10 and 20 across experiments.

5. Additional Experimental Details

In this section, we provide additional details about the TWDS16 dataset, including its quality, annotation process, and limitations. Furthermore, we outline additional quantitative and qualitative results. All experiments were conducted three times, and the average results are reported.

5.1. The TWDS16 Dataset

To generate transitional weather images in the TWDS16 dataset, we employed the Spurious Correlation Generator (SCG) to process pairs of input images captured under distinct weather conditions. Prior to this, we constructed the Weather Driving (WD) dataset, which contains images of five distinct weather conditions, sunny, cloudy, rainy, snowy, and foggy, while maintaining a consistent background across scenes. Cloudy images were sourced from Cityscapes [11], rainy images from RainCityscapes [14] & MultiWeatherCity [23], and foggy images from Foggy Cityscapes [27]. Snowy images are generated using ControlNet [39], a stable diffusion-based model. After addressing class imbalance, each discrete weather class in the WD dataset comprises 293 images, all resized to a resolution of 512 x 256 pixels. The final TWDS16 dataset consists of 46,880 weather transition sequences, each sampled uniformly at a length of T=10 frames. This dataset covers sixteen distinct weather transition states, namely, cloudy to rainy (CR), sunny to foggy (SF), sunny to rainy (SR), cloudy to snowy (CSn), cloudy to foggy (CF), snowy to rainy (SnR), snowy to foggy (SnF), foggy to rainy (FR) and vice versa. The total disk size of the TWDS16 dataset is 7 GB. Figures 4, 5, 6, and 7 present the qualitative results of the generated transitional weather sequences of TWDS16 dataset. The proposed SCG synthesizes smooth progressions between distinct weather conditions (e.g., cloudy, rainy, foggy, snowy, and sunny), demonstrating its ability to capture continuous and realistic visual transitions across a specified sequence length T.

Evaluating TWDS16 dataset quality. To ensure that the generated TWDS16 dataset aligns with real-world scenarios, we assess image quality using the Inception Score (IS), Signal-to-Noise Ratio (SNR), and Peak Signal-to-Noise Ratio (PSNR). Table 2 summarizes image quality results, where TWDS16 achieves higher IS, PSNR, and SNR scores, confirming its superior quality compared to existing benchmarks. Among the transitions, RC and FS yield the highest scores, while CR and SR deliver the next best performance.

Dataset annotations. We performed two types of image annotation to support both semantic and instance segmentation tasks. For semantic segmentation, we used the LabelImg tool to generate binary masks that distinguish between road and non-road regions. For instance, segmentation, we employed the Roboflow platform to annotate 12 object classes relevant to autonomous driving: car, truck, bus, bicycle, pedestrian, bike, obstacle, rider, tram, traffic light, van, and scooter.

Limitations of TWDS16 dataset. While the TWDS16

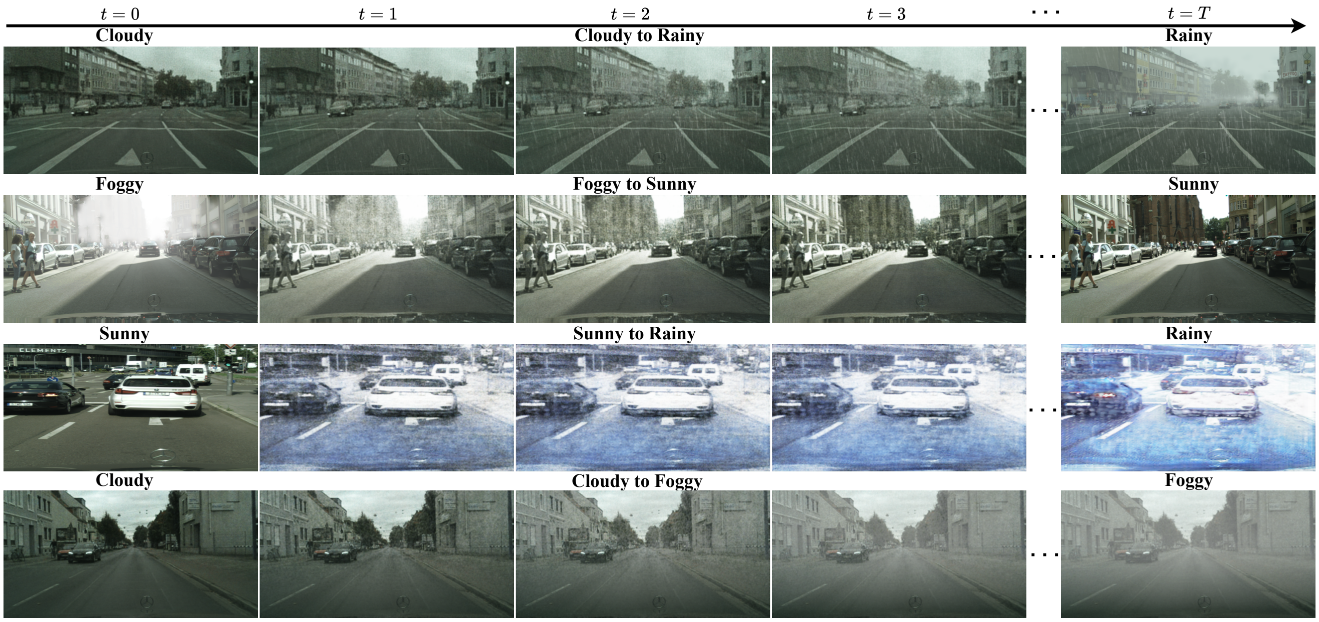


Figure 4. Qualitative results of TWDS16 dataset weather transitions (e.g., cloudy to rainy, foggy to sunny, sunny to rainy, and cloudy to foggy) generated by SCG. For a given length T, the model produces smooth and realistic transitions between weather conditions.

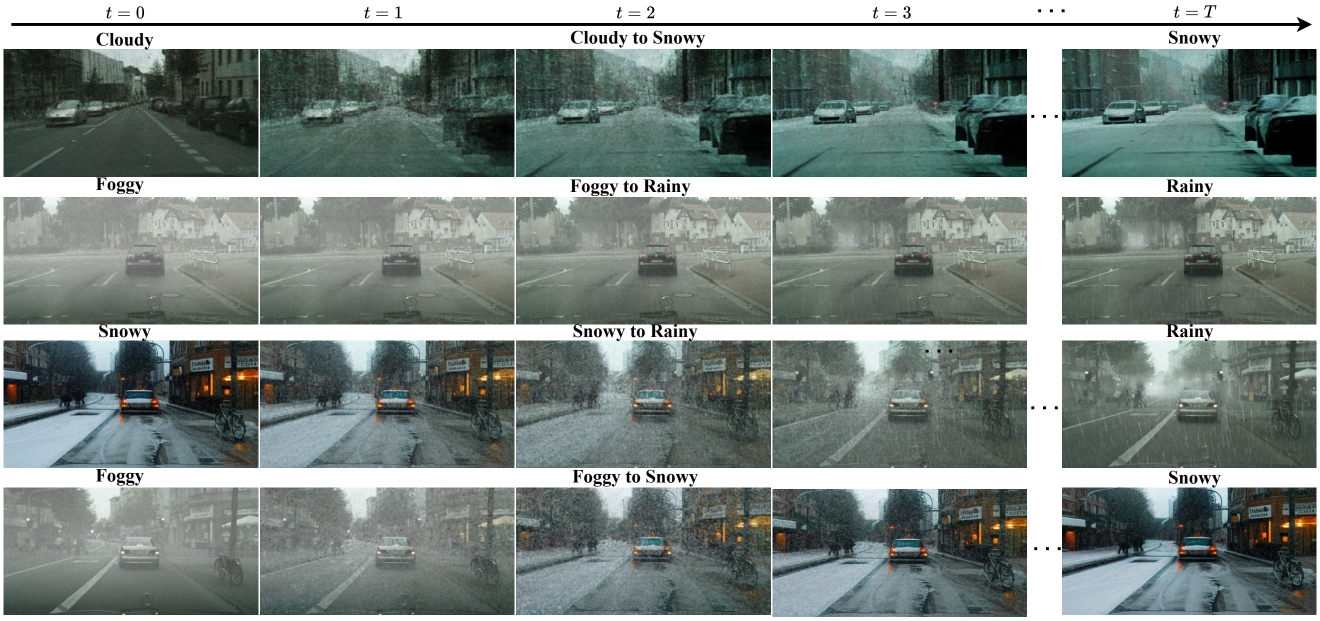


Figure 5. Qualitative results of TWDS16 dataset weather transitions (e.g., cloudy to snowy, foggy to rainy, snowy to rainy, and foggy to snowy) generated by SCG. For a given length T, the model produces smooth and realistic transitions between weather conditions.

dataset provides diverse intensities of weather conditions, it has certain limitations and potential biases, which we intend to investigate further in future work. *(i)* Limited geographic representation: The dataset contains static background scenes throughout a weather transition sequence, representing only a limited geographic area. Our future work aims to generate transition sequences with dynamic backgrounds and apply the CaRS method for segmentation. *(ii)* Temporal sampling bias: The dataset captures transitions at 10 frames per second, which may not adequately represent all weather intensity variations. This fixed sampling rate could introduce bias by missing intermediate

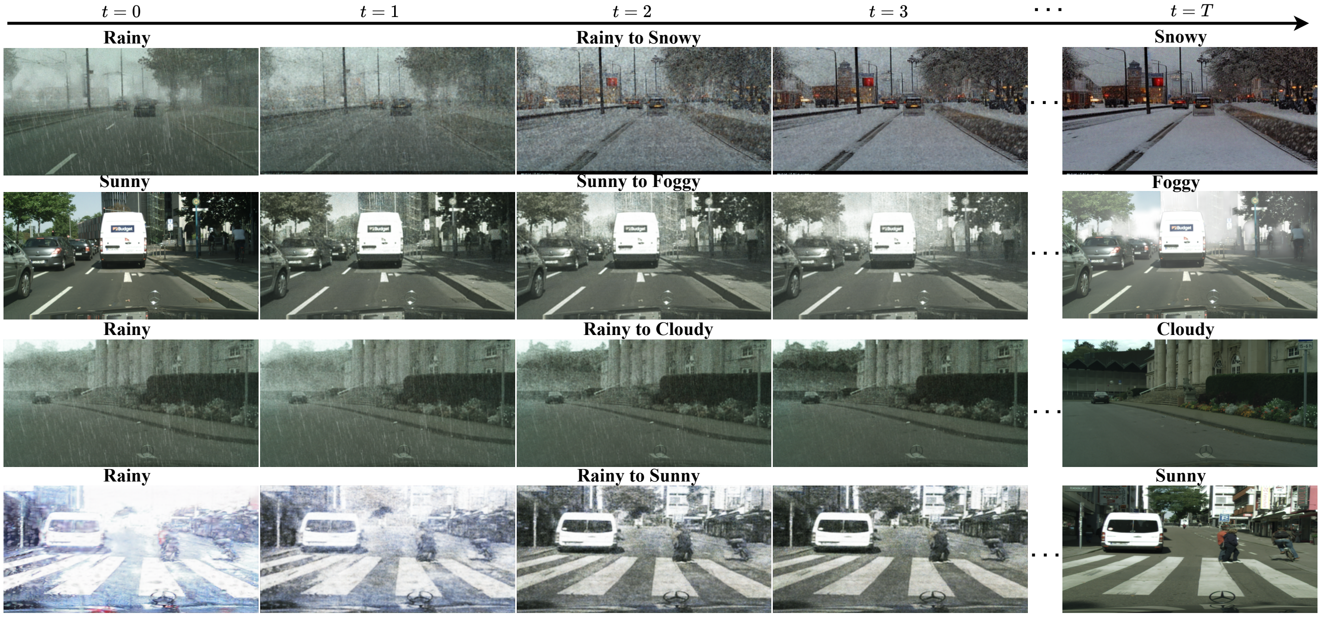


Figure 6. Qualitative results of TWDS16 dataset weather transitions (e.g., rainy to snowy, sunny to foggy, rainy to cloudy, and rainy to sunny) generated by SCG. For a given length T, the model produces smooth and realistic transitions between weather conditions.

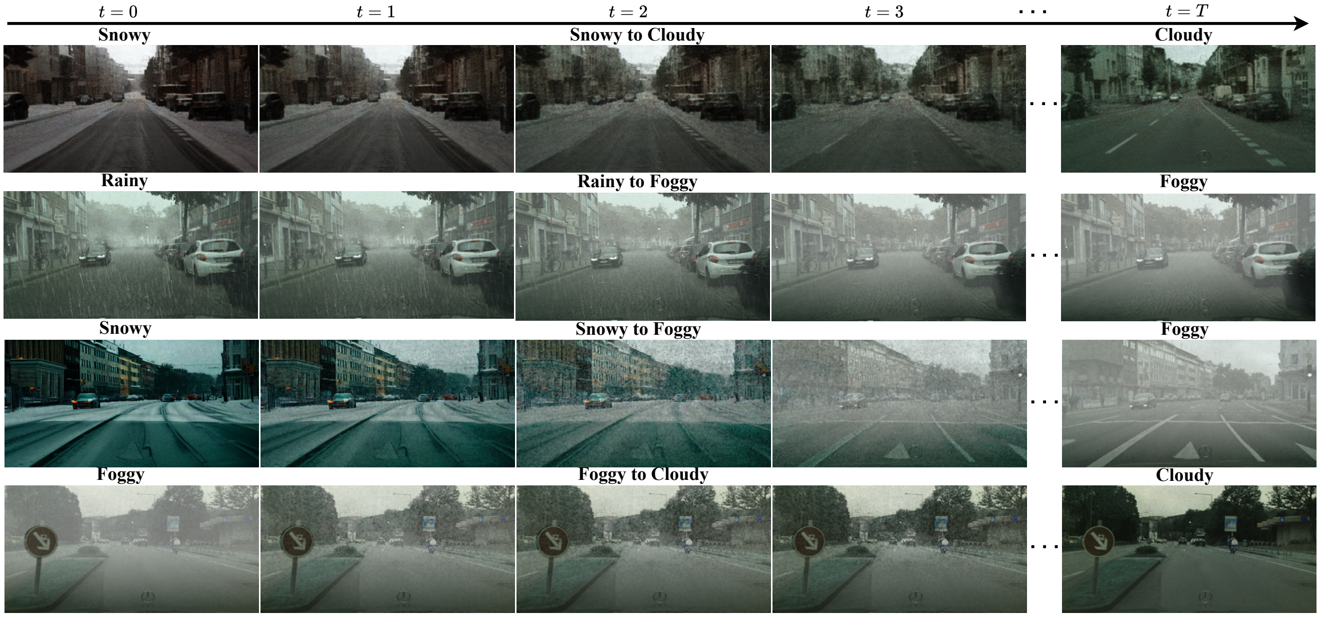


Figure 7. Qualitative results of TWDS16 dataset weather transitions (e.g., snowy to cloudy, rainy to foggy, snowy to foggy, and foggy to cloudy) generated by SCG. For a given length T, the model produces smooth and realistic transitions between weather conditions.

weather states. In future work, we will use different sampling rates and create annotations for them. *(iii)* Potential biases in data division: The dataset is divided randomly into training, validation, and testing sets. However, this approach may not account for biases in the data distribution, such as variations in the number of images of various weather scenarios, traffic scenarios or road elements. *(iv)* Annotation biases for road elements: The accuracy and consistency of the annotations for key road components, such as vehicles, pedestrians, traffic signs, and other critical infrastructure, could be subject to human biases or errors during the labeling process. This could lead to systematic under-

Table 2. Quality of TWDS16 images measured using IS, PSNR, and SNR metrics.

Dataset/ Metrics	BDD100K [36]	Cityscapes [11]	MW City [23]	CR	RC	SF	FS	SR	RS	CF	FC	CSn	SnC	SnR	RSn	SnF	FSn	FR	RF
IS ↑	4.51	5.00	4.13	54.70	**58.66**	49.39	54.91	58.28	48.17	50.44	51.11	49.95	48.55	46.55	45.34	48.34	47.55	52.47	51.92
PSNR ↑	-	27.73	31.59	32.59	30.10	31.07	**32.62**	31.22	31.54	31.42	31.15	30.52	30.33	30.11	30.32	29.76	30.01	31.77	31.54
SNR ↑	-	-	-	32.75	34.20	33.57	**34.25**	34.23	33.37	31.87	32.66	31.22	31.36	31.19	31.38	30.15	30.93	31.22	31.01

Table 3. Performance comparison of semantic road segmentation with existing models on the TWDS16 dataset, evaluated in terms of Dice score. **Bold** and result highlight the best and second best results, respectively.

Transition/ Model	RC	CR	SF	FS	SR	RS	CF	FC	CSn	SnC	SnR	RSn	SnF	FSn	FR	RF
DeepLabv3+ [8]	93.4	92.7	88.3	88.1	92.8	92.7	88.5	87.2	84.3	88.1	88.5	89.3	87.7	88.8	89.6	89.7
TransUNet [7]	93.1	93.5	87.0	86.1	93.4	92.5	79.3	79.5	78.5	79	87.5	88.7	80.5	82.6	82.8	82.4
Mask2Former [9]	91.2	92.8	87.6	87.2	90.5	91.8	91.2	90.3	88.8	87.2	88.7	90.2	91	90.1	91.2	89.9
SS-SFDA [16]	92.7	93.1	86.1	85.2	92.7	90.6	90.6	86.8	87.6	86.9	85.8	87.8	80.6	81.6	81.2	81
VBLC [18]	92.5	92.9	87.2	88.4	92.5	91.6	91.6	92.3	88.6	87.5	88.1	90.5	90.4	89.3	90.7	90.9
CaRS (ours)	**94.2**	**94.1**	**89.6**	**90.0**	**94.7**	**93.9**	**95.3**	**94.6**	**91.1**	**92**	**92.9**	**93.8**	**91.6**	**91**	**93.6**	**93.2**

representation or misclassification of certain road elements, which may impact the model's ability to generalize to real-world scenarios. Evaluating the quality and potential biases in the road element annotations should be a focus of future investigations.

In addition to our generated TWDS16 dataset, we evaluated the proposed method on Foggy Cityscapes [27], RainCityscapes [14], and BDD100K [36], with comparative results presented in the next section.

5.2. Additional Quantitative and Qualitative Results

Table 4 in the main paper reports the semantic segmentation performance of CaRS against baselines using mIoU. Building on this, Tables 3 and 4 present results in terms of Dice score and accuracy across all weather transitions in the TWDS16 dataset. CaRS consistently outperforms state-of-the-art semantic segmentation models (DeepLabV3+, Mask2Former) and domain-adaptive methods (SS-SFDA, VBLC) designed for adverse weather. Notably, the CF and FC transitions yield the highest gains, underscoring CaRS's robustness under challenging shifts. While DeepLabV3+ and Mask2Former achieve competitive accuracy, they fall short of CaRS's overall performance.

Table 5 of the main paper reports the instance segmentation performance of CaRS against baselines using mask-mAP. Additionally, Table 5 presents a quantitative evaluation in terms of box-mAP across all transition states of the TWDS16 dataset. In both evaluations, CaRS consistently outperforms all baseline methods. Figure 8 illustrates the instance segmentation performance of CaRS across various weather transitions, RC, CF, FS, FR, RS, and CSn, evaluated at different weather intensity levels (t). The results show that CaRS effectively mitigates the confounding effects introduced by varying weather patterns, maintaining high segmentation accuracy even under severe weather transitions. In contrast, baseline models exhibit a significant drop in accuracy as the intensity of the weather transition increases.

Figure 9 shows the test losses for semantic road segmentation across the transitions CR, FC, SF, FSn, SR, and RSn, comparing CaRS with baseline models. CaRS consistently achieves the lowest and most stable test loss among all methods. Similarly, Figure 10 shows test losses for instance segmentation of road elements across the same set of transitions, where CaRS again outperforms others with consistently lower losses. However, for SR and RSn transitions, the test loss of CaRS is relatively higher compared to other transitions, indicating potential for further improvement in future work. Furthermore, we generate Grad-CAM heatmaps to visually demonstrate the effectiveness of our proposed approach. Figure 11 shows Grad-CAM heatmaps before and after causal–confounding disentanglement. Without disentanglement, attention maps are scattered across irrelevant regions, reducing the ability to localize true causal features because of spurious weather correlations. After disentanglement, attention becomes more structured and aligned with the target classes: road heatmaps focus on road areas, car heatmaps correctly localize all car instances, and person heatmaps concentrate on pedestrian regions. This demonstrates that the proposed disentanglement framework suppresses spurious correlations and enhances causal feature learning, resulting in improved segmentation performance.

5.2.1. Average Causal Effect (ACE)

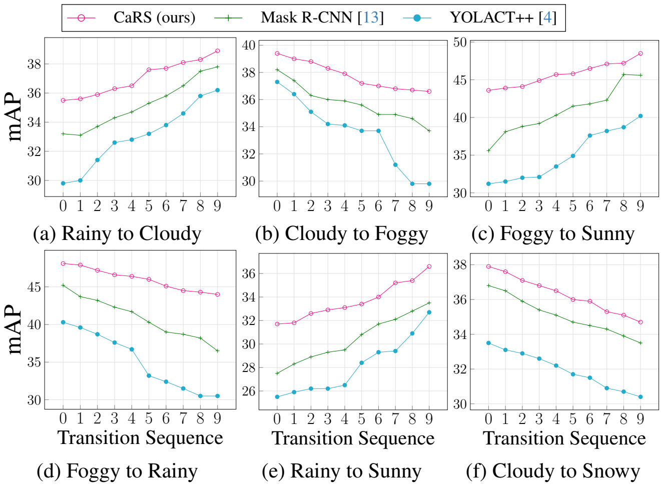
We computed the ACE values for semantic road segmentation and road element instance segmentation under the RC

Table 4. Accuracy of semantic road segmentation with existing models on the TWDS16 dataset. **Bold** and result highlight the best and second best results, respectively.

Transition/ Model	**RC**	**CR**	**SF**	**FS**	**SR**	**RS**	**CF**	**FC**	**CSn**	**SnC**	**SnR**	**RSn**	**SnF**	**FSn**	**FR**	**RF**
DeepLabv3+ [8]	90.4	91.9	93.1	93.2	94.5	94.3	92.8	91.7	92.5	92.2	92	91.9	92.1	92	92.5	92.3
TransUNet [7]	91.0	91.1	90.8	91.7	94.1	95.2	92.6	93.6	88.7	88.2	86.3	87.2	86.2	85.8	88.5	87.9
Mask2Former [9]	91.5	91.7	92.9	91.5	92.2	94.6	93.1	93.9	92.2	91.8	91.9	92.5	91.6	91.6	91.8	91.7
SS-SFDA [16]	89.2	88.8	88.2	90.6	92.4	91.3	92.2	91.3	90.1	89.7	85.4	85.9	84.5	84.9	85.2	84.3
VBLC [18]	91.0	91.1	90.5	92.5	93.1	93.9	93.9	93.5	91.7	92.1	91.1	91.8	91.2	92.2	91.8	91.9
CaRS (ours)	**92.1**	**92.2**	**93.7**	**93.7**	**94.2**	**95.1**	**95.9**	**95.8**	**93**	**93.1**	**93.3**	**94.3**	**92.2**	**93.4**	**93.9**	**93.4**

Table 5. Performance comparison between the proposed method and state-of-the-art models for road element instance segmentation on the TWDS16 dataset, evaluated in terms of bounding box mAP. **Bold** and result highlight the best and second best results, respectively.

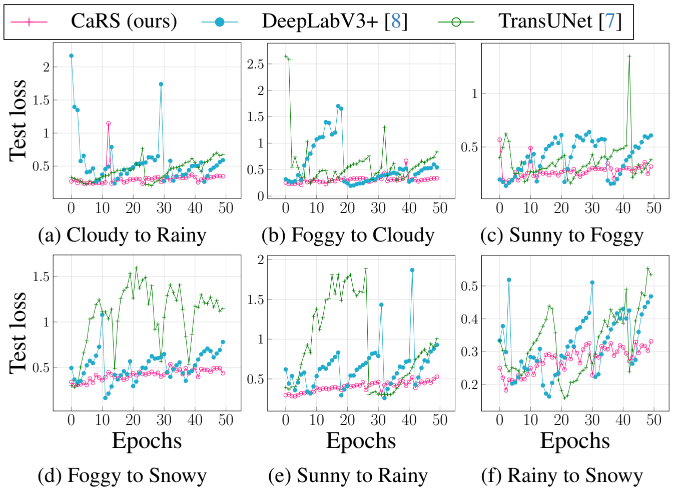
Transition/ Model	**RC**	**CR**	**SF**	**FS**	**SR**	**RS**	**CF**	**FC**	**CSn**	**SnC**	**SnR**	**RSn**	**SnF**	**FSn**	**FR**	**RF**
Mask R-CNN [13]	36.3	36.9	38.5	39.9	29.6	30.6	39.2	39.8	36	35.2	30.1	29.9	32.9	33	36.7	35.9
YOLACT++ [4]	31.8	31.3	32.6	31.7	29.6	28.9	33.3	33.7	30.5	30.3	27.8	28.3	27.6	28.8	29.7	30.8
SparseInst [10]	35.8	36.4	38.5	40.6	32.5	30.4	38.9	37.5	35.4	35.8	31.5	31.0	31.5	30.3	37.5	37.4
Mask2Former [9]	36.4	36.7	34.6	36.0	29.5	30.9	38.6	38.4	36.2	36.3	31.8	31.2	31.4	31.9	36.3	37.5
CaRS (ours)	**37.5**	**37.9**	**45.2**	**45.9**	**35.2**	**33.5**	**44.7**	**44.4**	**37.2**	**37.5**	**34.2**	**33.9**	**35.5**	**36.7**	**41.2**	**40.5**



(a) Rainy to Cloudy (b) Cloudy to Foggy (c) Foggy to Sunny

(d) Foggy to Rainy (e) Rainy to Sunny (f) Cloudy to Snowy

Figure 8. Comparing the instance segmentation performance of the CaRS method with the state-of-the-art models on the TWDS16 dataset across all the intensities of the weather.



(a) Cloudy to Rainy (b) Foggy to Cloudy (c) Sunny to Foggy

(d) Foggy to Snowy (e) Sunny to Rainy (f) Rainy to Snowy

Figure 9. Comparison of semantic road segmentation loss of the CaRS with state-of-the-art models on the TWDS16 dataset.

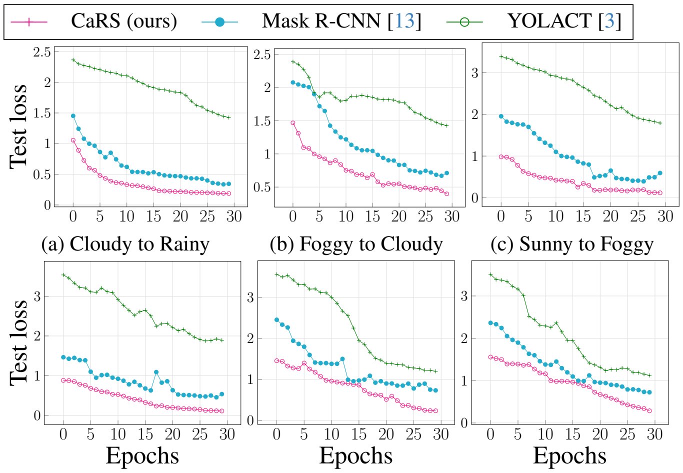
weather transition, across intensity levels from $t = 0$ to 9. These values, calculated using Eq.2 from the main paper and visualized in Figure 12, reveal a consistent decline in ACE as weather intensity decreases. CaRS exhibits significantly lower ACE values than other methods, indicating its superior ability to mitigate weather-induced biases. This suggests that CaRS effectively addresses confounding factors, leading to more robust and reliable segmentation performance across varying weather intensities.

5.3. Extended Ablation Study

In continuation from the main paper, we present additional ablation studies evaluating the proposed CaRS method on Foggy Cityscapes [27] (5,500 images) and RainCityscapes [14] (15,000 images). The main paper reported quantitative results on these benchmarks using dataset splits consis-

Table 6. Semantic road segmentation performance of the CaRS method on unseen weather transitions of the TWDS16 dataset (Trained on CR and RS transitions). **Bold** and result highlight the best and second best results, respectively. mI→mIOU, DI→Dice score.

Test/ Train		RC	CR	SF	FS	SR	RS	CF	FC	SnR	RSn	SnF	FSn	CS	SC	FR	RF
CR	mI	0.48	-	0.44	0.43	0.37	0.58	0.6	**0.59**	0.45	0.44	0.46	0.48	0.39	0.42	0.49	0.48
	DI	0.65	-	0.6	0.6	51.6	0.74	0.77	0.75	0.61	0.60	0.62	**0.66**	0.54	0.59	**0.66**	0.64
RC	mI	0.47	**0.6**	0.45	0.48	0.5	-	0.46	0.44	0.46	0.49	0.47	0.5	0.44	0.46	0.52	0.5
	DI	0.64	0.75	0.63	0.66	0.67	-	**0.68**	0.65	0.62	0.64	0.63	0.65	0.61	0.62	**0.68**	0.65



(a) Cloudy to Rainy (b) Foggy to Cloudy (c) Sunny to Foggy

(d) Foggy to Snowy (e) Sunny to Rainy (f) Rainy to Snowy

Figure 10. Comparison of instance segmentation loss of the CaRS with state-of-the-art models on the TWDS16 dataset.

Table 7. Performance of CaRS on unseen Foggy Cityscapes (FC), Rainy Cityscapes (RC), and real-world adverse weather scenarios from BDD100K, when trained on the TWDS16 dataset. mI→mIOU, DI→Dice Score.

Dataset	**Semantic Segmentation**		**Instance Segmentation**	
	DI	**mI**	**box**	**seg**
RC	0.67	0.5	0.31	0.28
FC	0.63	0.46	0.32	0.29
BDD	0.58	0.44	0.29	0.27

tent with TWDS16. For comparison, we included multiple baselines: DeepLabv3+, TransUNet, and U-Net for semantic segmentation, and Mask R-CNN and YOLACT++ for instance segmentation. We further evaluated instance segmentation on BDD100K [36], which contains 100,000 driving video clips. From this, 700 clips were used for training, 100 for validation, and 200 for testing. Comparative quantitative results are provided in the main paper. Figures 13 and 14 show qualitative results on Foggy Cityscapes and RainCityscapes, where CaRS effectively segments objects despite challenging weather. Similarly, Figure 15 demonstrates robust segmentation of road elements across clear, adverse weather, and nighttime conditions in BDD100K.

In terms of computational efficiency, CaRS offers significant advantages. Despite being trained on fewer samples, it achieves competitive performance while requiring fewer parameters and floating-point operations (see Tables 3 and 7 in the main paper). Moreover, it attains lower inference time, FPS, and memory usage compared to other models. The complete pipeline, including SCG generation of the TWDS16 dataset (1,050 s), semantic road segmentation (3,000 s), and instance segmentation of road elements (5,430 s), further demonstrates its reduced overall computational complexity.

Generalizability of the Proposed approach. We evaluated the generalization ability of CaRS in unseen weather environments. In the first, CaRS is trained on the CR transition and tested across all other transitions in the TWDS16 dataset. In the second, CaRS training is performed on the RS transition, and testing was carried out again on the remaining transitions. As shown in Table 6, CaRS achieved strong segmentation performance, even when tested on weather transitions not seen during training. Furthermore, Table 7 reports the performance of CaRS trained on TWDS16 and evaluated on three external datasets, FoggyCityscapes, RainCityscapes, and the real-world BDD100K dataset, demonstrating its robustness under real-world adverse weather beyond the source domain. Additionally, the dual attention mechanism used to separate causal and confounding features is model-agnostic, allowing seamless integration with diverse backbone architectures.

5.4. Discussion on Edge Cases

Unusual weather patterns. To enhance model robustness across diverse weather conditions, our training data incorporates transitional weather patterns with varying intensity levels, ranging from $t = 0$ to T (e.g., $T = 10$). As discussed in Section 3.1 of the main paper, the VAE performs data interpolation using the parameter $\alpha \in [0, 1]$, which controls the intensity level of weather conditions. As a result, the generated dataset captures a diverse range of weather patterns. Although this allows the model to adapt effectively

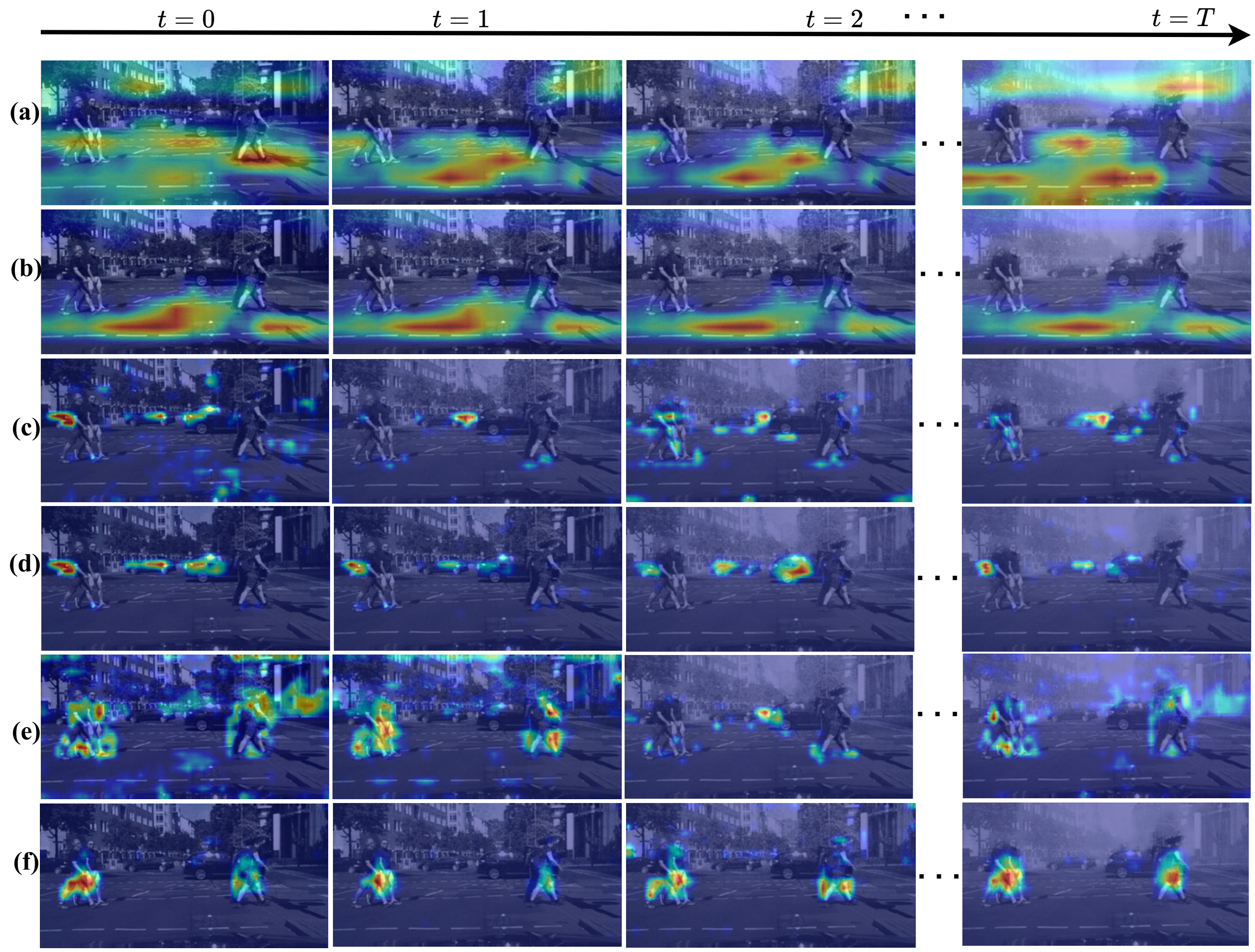
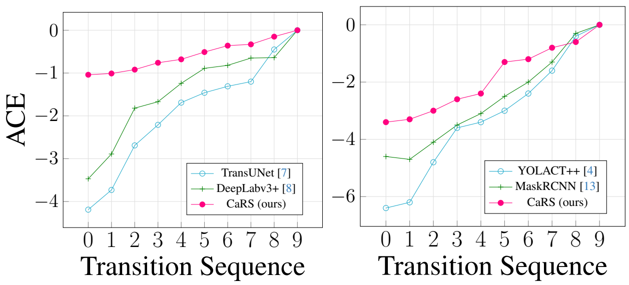


Figure 11. Grad-CAM heatmaps before and after causal–confounding disentanglement, illustrating improved focus on causal regions across the transition sequence. (a) Road segmentation before disentanglement, where attention spreads across irrelevant regions. (b) Road segmentation after disentanglement, with attention correctly focused on road areas. (c) Car segmentation before disentanglement, with diffuse attention beyond car objects. (d) Car segmentation after disentanglement, with precise focus on all car instances. (e) Person segmentation before disentanglement, where attention is scattered across the image. (f) Person segmentation after disentanglement, with accurate focus on all person instances.



(a) Semantic segmentation

(b) Instance segmentation

Figure 12. Average causal effect values of rainy to cloudy transition of the TWDS16 dataset. (a) Semantic road segmentation (b) Instance segmentation of road elements.

and maintain strong performance in many scenarios, further improvements are needed to ensure optimal performance in unusual weather patterns.

Performance on real-world noise and adversarial robustness. Our model performs well across challenging weather conditions, but real-world noise and adversarial perturbations may introduce additional challenges. Weather transitions often exhibit irregular patterns, where certain dependencies are more critical than others. To further strengthen robustness, we plan to explore adversarial defense strategies, including adversarial training and uncertainty quantification techniques, to mitigate potential vulnerabilities. These enhancements will ensure our model maintains strong generalization across real-world weather

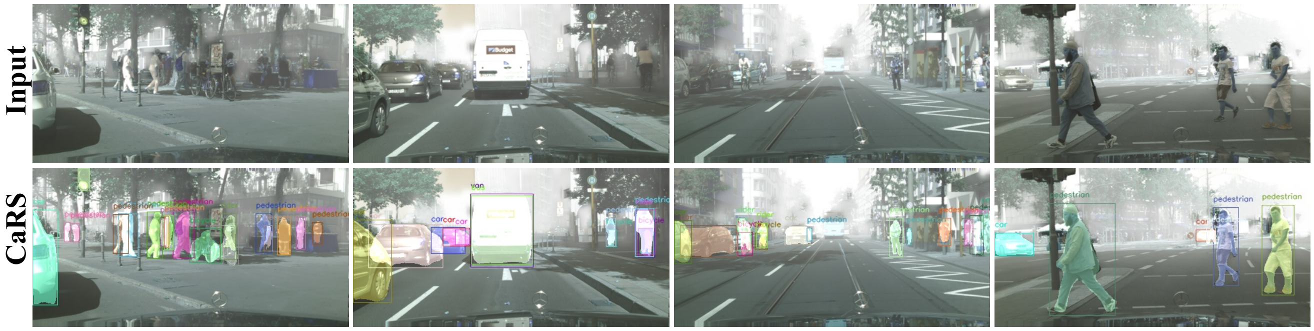


Figure 13. The qualitative instance segmentation results of the CaRS method on the Foggy Cityscapes dataset.

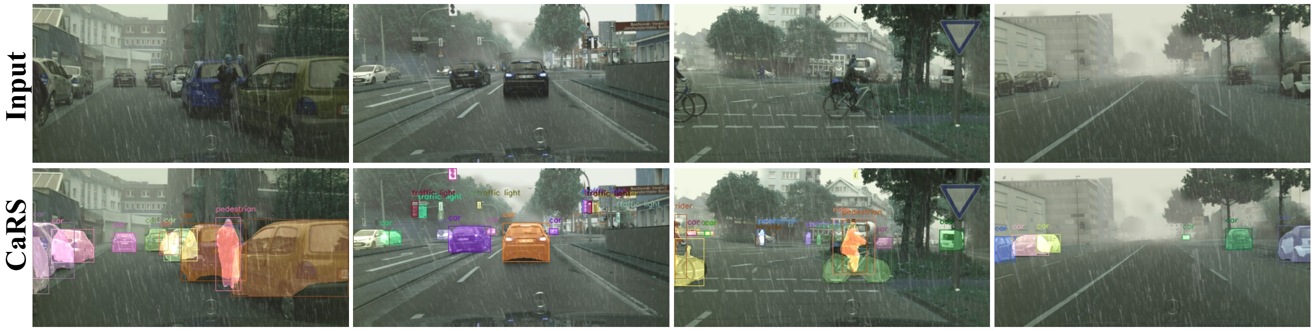


Figure 14. The qualitative instance segmentation results of the CaRS method on the RainCityscapes dataset.

conditions.

6. Limitations and Future Work

In continuation of Section 5 of the main paper, CaRS has certain limitations that we plan to address in future research.

1. **Background consistency.** The current weather transition generation maintains static background elements across transitions, modifying only weather-related features. Although this controlled setup simplifies evaluation by reducing weather effects on the model performance, it limits real-world applicability. The challenge lies in generating dynamic backgrounds that change naturally while preserving accurate weather transitions, requiring a balance between scene variation and weather feature clarity. Future research will focus on transitional video generation that integrates dynamic backgrounds while maintaining precise weather details, demanding advanced architectures to effectively separate weather and scene features.
2. **Transition quality.** Image quality deteriorates during transitions from sunny to rainy, rainy to sunny, snowy to rainy and rainy to snowy due to high distortion and noise in rainy and snowy images. To address this challenge, future work will incorporate denoising algorithms as a preprocessing step to mitigate visual noise and distortions, ensuring cleaner inputs for transition generation. This enhancement will ultimately improve CaRS's ability to handle challenging weather shifts.
3. **Unusual weather patterns.** To enhance CaRS's robustness across diverse weather conditions, our generated data includes transitional weather patterns with varying intensity levels, ranging from $t = 0$ to T. While this enables the model to adapt effectively and maintain strong performance in many scenarios, further improvements can be achieved through unsupervised domain adaptation to perform well in unseen weather environments.
4. **Model robustness to adversarial perturbations.** While the dataset captures transitional weather conditions, high noise levels in the images may affect model robustness, reducing reliability in real-world scenarios, particularly under natural noise or adversarial perturbations. To address this, future work will explore adversarial attack and defense strategies to enhance the CaRS's resilience in real-world deployments.
5. **Risk analysis.** While our work focuses on robust segmentation under transitional weather, a promising direction is to integrate risk analysis into the segmentation pipeline. By localizing potential hazards, future models can move beyond semantic understanding to reason about risks in real time. Such risk-aware perception is vital for safe autonomous driving, especially in transitional weather conditions [21, 40]. Coupling segmentation and risk localization with contextual explanation en-

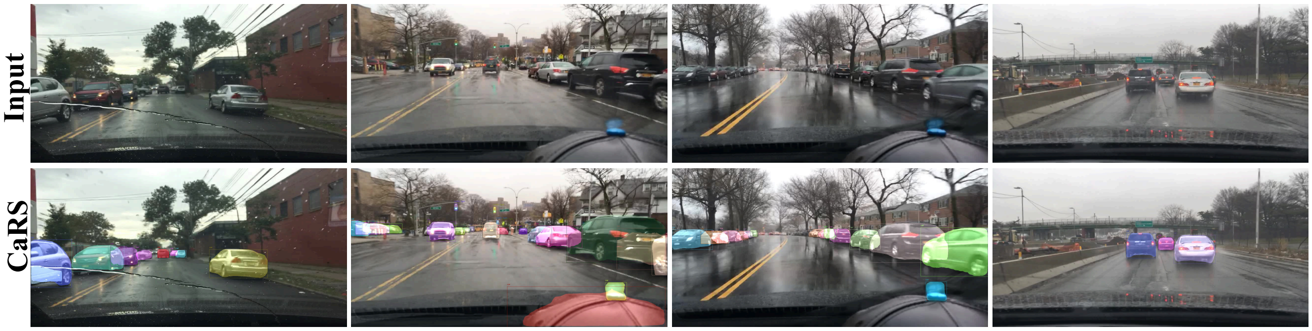


Figure 15. The qualitative instance segmentation results of the CaRS method on the BDD100K dataset.

ables more informed, interpretable decision-making. Incorporating causal reasoning into risk assessment (e.g., distinguishing true danger signals from spurious noise) can further enhance safety, generalization, and trust in autonomous systems.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 4

[2] Tariq Berrada, Camille Couprie, Karteek Alahari, and Jakob Verbeek. Guided distillation for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 475–483, 2024. 1

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 10

[4] David Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2020. 9, 11

[5] David Brüggemann, Christos Sakaridis, Tim Brödermann, and Luc Van Gool. Contrastive model adaptation for cross-condition robustness in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11378–11387, 2023. 1

[6] Yingfeng Cai, Lei Dai, Hai Wang, and Zhixiong Li. Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving. *IEEE Transactions on Image Processing*, pages 9069–9084, 2021. 1

[7] Haofeng Chen, Qi Dou Zhang, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Transunet: Transformers make strong encoders for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 407–418. Springer, 2021. 8, 9, 11

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 8, 9, 11

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 4, 8, 9

[10] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022. 1, 4, 9

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5, 8

[12] Gianni Franchi, Marwane Hariat, Xuanlong Yu, Nacim Belkhir, Antoine Manzanera, and David Filliat. Infraparis: A multi-modal and multi-task autonomous driving dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2973–2983, 2024. 1

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2961–2969, 2017. 9, 10, 11

[14] Haofeng Hu, Liang Chen, Zhongyuan Wang, Lei Zhu, Mengyang Zhang, and Yun Fu. Depth-attentional features for single-image rain removal. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8022–8031, 2019. 5, 8, 9

[15] Jie Hu, Yao Lu, Shengchuan Zhang, and Liujuan Cao. Istr: Mask-embedding-based instance segmentation transformer. *IEEE Transactions on Image Processing*, 33:2895–2907, 2024. 1

[16] Divya Kothandaraman, Rohan Chandra, and Dinesh Manocha. Ss-sfda: Self-supervised source-free domain adaptation for road segmentation in hazardous environments. In *Proceedings of the IEEE/CVF International Conference*

on *Computer Vision Workshops*, pages 3049–3059, 2021. 1, 8, 9

[17] Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18911–18921, 2022. 1

[18] Mingjia Li, Binhui Xie, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Vblc: Visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8605–8613, 2023. 1, 8, 9

[19] Wei Li and Zhixin Li. Causal-setr: a segmentation transformer variant based on causal intervention. In *Proceedings of the Asian conference on computer vision*, pages 756–772, 2022. 2

[20] Liang Liao, Wenyi Chen, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Unsupervised foggy scene understanding via self spatial-temporal label diffusion. *IEEE Transactions on Image Processing*, pages 3525–40, 2022. 1

[21] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052, 2023. 12

[22] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21426–21437, 2023. 2

[23] Valentina Musat, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. Multi-weather city: Adverse weather stacking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2915, 2021. 1, 5, 8

[24] Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M Schmidt, and Tim 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*, pages 2830–2840, 2024. 1

[25] Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14874–14884, 2024. 2

[26] Giulia Rizzoli, Matteo Caligiuri, Donald Shenaj, Francesco Barbato, and Pietro Zanuttigh. When cars meet drones: Hyperbolic federated learning for source-free domain adaptation in adverse weather. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1587–1596, 2025. 1

[27] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5, 8, 9

[28] Furqan Ahmed Shaik, Abhishek Reddy, Nikhil Reddy Billa, Kunal Chaudhary, Sunny Manchanda, and Girish Varma. Idd-aw: A benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4614–4623, 2024. 1

[29] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1417–1427, 2021. 5

[30] Rahul Venkataramani, Parag Dutta, Vikram Melapudi, and Ambedkar Dukkipati. Causal feature alignment: Learning to ignore spurious background features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4666–4674, 2024. 2

[31] Kaihong Wang, Donghyun Kim, Rogerio Feris, and Margrit Betke. Cdac: Cross-domain attention consistency in transformer for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11519–11529, 2023. 1

[32] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 2

[33] Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099, 2023. 1

[34] Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8103–8112, 2023. 2

[35] Chengxiang Yin, Jian Tang, Tongtong Yuan, Zhiyuan Xu, and Yanzhi Wang. Bridging the gap between semantic segmentation and instance segmentation. *IEEE Transactions on Multimedia*, 24:4183–4196, 2022. 1

[36] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *The International Journal of Robotics Research*, 39(1):1–18, 2020. 8, 10

[37] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in neural information processing systems*, 33:655–666, 2020. 2

[38] Hua Zhang, Liqiang Xiao, Xiaochun Cao, and Hassan Foroosh. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1742–1756, 2024. 2

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[40] Yue Zhang, Yajie Zou, Yunlong Zhang, Lingtao Wu, et al. Spatiotemporal interaction pattern recognition and risk evolution analysis during lane changes. *IEEE Transactions on Intelligent Transportation Systems*, 24(6):6663–6673, 2023. 12

[41] Zhiyong Zheng, Xu Li, Qimin Xu, and Xiang Song. Deep inference networks for reliable vehicle lateral position estimation in congested urban environments. *IEEE transactions on image processing*, pages 8368–8383, 2021. 1

[42] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13760–13769, 2022. 1

[43] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Xuequan Lu, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8954–8968, 2023. 1