Parsing All Adverse Scenes: Severity-Aware Semantic Segmentation with Mask-Enhanced Cross-Domain Consistency

Fuhao Li[1*], Ziyang Gong[2*], Yupeng Deng[3*], Xianzheng Ma[4†], Renrui Zhang[5], Zhenming Ji[2], Xiangwei Zhu[2], Hong Zhang[1†]

[1]Wuhan University of Science and Technology
[2]Sun Yat-Sen University
[3]National University of Singapore
[4]University of Oxford
[5]Shanghai AI Lab

Abstract

Although recent methods in Unsupervised Domain Adaptation (UDA) have achieved success in segmenting rainy or snowy scenes by improving consistency, they face limitations when dealing with more challenging scenarios like foggy and night scenes. We argue that these prior methods excessively focus on weather-specific features in adverse scenes, which exacerbates the existing domain gaps. To address this issue, we propose a new metric to evaluate the severity of all adverse scenes and offer a novel perspective that enables task unification across all adverse scenarios. Our method focuses on Severity, allowing our model to learn more consistent features and facilitate domain distribution alignment, thereby alleviating domain gaps. Unlike the vague descriptions of consistency in previous methods, we introduce Cross-domain Consistency, which is quantified using the Structural Similarity Index Measure (SSIM) to measure the distance between the source and target domains. Specifically, our unified model consists of two key modules: the Merging Style Augmentation Module (MSA) and the Severity Perception Mask Module (SPM). The MSA module transforms all adverse scenes into augmented scenes, effectively eliminating weather-specific features and enhancing Cross-domain Consistency. The SPM module incorporates a Severity Perception mechanism, guiding a Mask operation that enables our model to learn highly consistent features from the augmented scenes. Our unified framework, named PASS (Parsing All adverSe Scenes), achieves significant performance improvements over state-of-the-art methods on widely-used benchmarks for all adverse scenes. Notably, the performance of PASS is superior to Semi-Unified models and even surpasses weather-specific models.

Introduction

Adverse scene understanding is always challenging for autonomous driving perception due to unpredictable noise information in severe observations. Additionally, the large discrepancy between different adverse scenes, such as fog, night, rain, and snow, makes it particularly difficult to solve all of them with a unified model. While recent methods have attempted to address adverse scene tasks by improving consistency (Hoyer et al. 2023), they still have limitations.

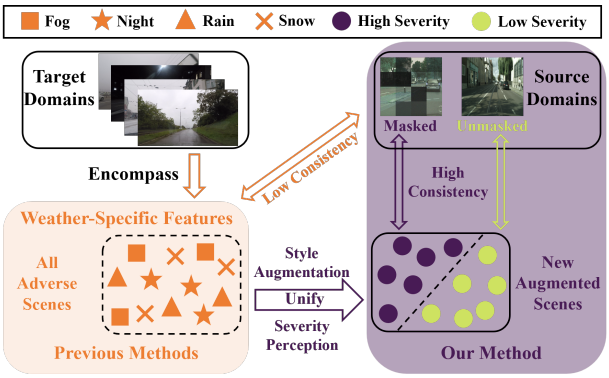

Figure 1: Previous methods excessively focus on the weather-specific features of all adverse scenes causing their weaknesses. In contrast, our method eliminates these weather-specific features by transforming them into new augmented scenes. Then our method will be guided by Severity Perception to divide augmented images into images with high and low severity corresponding to whether to mask source domain images to strengthen Cross-domain Consistency. We note the solid box means the images and the dashed box means the features.

Although prior works (Wang et al. 2022; Wang, Zhu, and Yang 2021) have achieved success in addressing rainy or snowy scenes by enhancing consistency, they show weaknesses when faced with more demanding scenarios like foggy and night scenes. We argue that previous methods (Wang et al. 2020) excessively focus on weather-specific features such as brightness, hue, and contrast conditions during the adaptation process, resulting in models overly sensitive to these specific features and lacking generalization across all adverse scenes.

To overcome this challenge, we propose that models should not prioritize learning weather-specific features inherent in all adverse scenes. Instead, we suggest unifying all adverse scenes by eliminating these intrinsic fea-

*These authors contributed equally.
†Corresponding authors.

tures to reduce model sensitivities and introducing a new metric to evaluate the severity of adverse scenes. By using severity as a guiding principle, our model can learn weather-agnostic features with high Cross-domain Consistency. We define Cross-domain Consistency as the measure of the distance between the source and target domains and argue that enhancing it can effectively reduce the disparity between domains, improve knowledge transfer efficiency, and enable models to learn more consistent features. Consequently, our method aims to enhance Cross-domain Consistency through two modules: Merging Style Augmentation (MSA) and Severity Perception Mask (SPM).

To eliminate weather-specific features, the MSA module operates on the target domains under adverse scenes, utilizing Style Augmentation (SA) and subsequent Image Merging. SA effectively eliminates weather-specific features by transforming all adverse scenes into new augmented scenes. We argue that SA enhances Cross-domain Consistency, as the augmented scenes without the influence of weather contain more consistent and aligned features with the source domain. However, we are concerned that SA may result in the loss of certain content information due to strong style transformations. To address this, we introduce an Image Merging mechanism within MSA, which merges augmented and original images to preserve content information. This is important as our baseline (Hoyer, Dai, and Van Gool 2022b) employs a high-resolution approach that is highly sensitive to content information compared to other methods.

By eliminating weather from the target domains, we aim to further decrease the sensitivity of our model to weather from the source domain. To achieve this, we introduce the SPM module, consisting of a Mask Operation and Severity Perception. Our Mask Operation differs from previous fully covered masks and focuses on adjusting the brightness, hue, contrast, and noise of images by randomly sampling from preset uniform distributions (details discussed later). This forces our model to learn to ignore weather-specific features from both the source and target domains. However, relying on the single Mask Operation has a detrimental effect on consistency, as masked images with high severity may not align well with all augmented scene images with varying severity levels. To alleviate this issue, we introduce the Severity Perception mechanism, which screens out high-severity and low-severity images and aligns them with masked and unmasked images, respectively. This enhances cross-domain consistency, as depicted in Fig 1.

In our experiments, we will conduct comprehensive investigations into Cross-domain Consistency, accompanied by sufficient ablation studies, to validate our intuitions. Our main contributions can be summarized as follows: (1) We are the first to introduce SSIM (Wang et al. 2004) to UDA for semantic segmentation, providing a quantifiable measure of Cross-domain Consistency. (2) We propose a unified model that can effectively handle tasks in all adverse scenes, including fog, night, rain, and snow. (3) Our method surpasses previous state-of-the-art approaches on widely-used benchmarks for various adverse scenes, such as ACDC (fog, night, rain, snow), Foggy Zurich, Foggy Driving, Dark Zurich, Nighttime Driving, and BDD100k-Night. Notably, our performance improvement on Cityscapes to Foggy Zurich exceeds 9% compared to the state-of-the-art methods.

Related Work

UDA for Semantic Segmentation To address the challenges generated by domain shift, many UDA methods have been proposed that focus on adversarial training (Tsai et al. 2018; Zhu et al. 2017; Hoffman et al. 2018; Liao et al. 2022) or self-training (Gong et al. 2023). In adversarial training, models promoting domain-invariant features always develop a discriminator to combine with a generative adversarial network (GAN) (Goodfellow et al. 2020). Self-training models leverage a teacher network to generate pseudo-labels (Lee et al. 2013) to conduct unsupervised learning.

Adverse Scenes Understanding In Unsupervised Domain Adaptation, previous works focusing on adverse scene understanding can be divided into foggy scene understanding, night scene understanding, and other methods (Brüggemann et al. 2023; Gao et al. 2022; Lei et al. 2020) focusing on several scenes. For the foggy scene understanding (Iqbal, Hafiz, and Ali 2022), previous works (Dai et al. 2020; Sakaridis et al. 2018; Sakaridis, Dai, and Van Gool 2018) have tried to generate synthetic foggy images as training data by developing fog simulators to transform the clear weather. However, the synthetic-to-real gaps still exist. Thus, there are two kinds of works to close the weather gaps (Ma et al. 2022) and focus on Image Dehazing (Lee, Son, and Kwak 2022). For the night scene understanding, previous night-specialized works (Xie et al. 2023) also focus on narrowing the gaps by introducing twilight images (Dai and Van Gool 2018), and extra geometry information to further refine predictions (Sakaridis, Dai, and Gool 2019). Notably, our unified model, PASS, can solve all adverse scene tasks and even outperforms these weather-specialized models.

All in One All-in-One Network (Li, Tan, and Cheong 2020; Valanarasu, Yasarla, and Patel 2022) was proposed to solve Imgae Restoration tasks under adverse scenes by handling multiple adverse scene degradations using a single network. Although the scope of research is different, the aim of PASS to solve all adverse scene tasks by a unified model is similar to All-in-One models.

Method

Overview of Our Method

The training datasets of models consist of source domain images $X_S \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ with GT label $Y_S \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ and unlabeled target domain images $X_T \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 3}$, where H, W, C respectively means that the height, width, and category of classes of images. The MSA will work on X_T to transform them into $X_{T'}$ and then merge it with X_T to final augmented scene images X_M. The SPM will transform X_S into $X_{S'}$ according to whether X_M passes the Severity Perception mechanism.

Merging Style Augmentation (MSA)

MSA focuses on transforming all adverse scenes into new augmented scenes that possess higher Cross-domain Consistency containing more consistent features. MSA consists

of two distinct components: Style Augmentation (SA) and Image Merging. To a better understanding of MSA, we need to recap the SA.

Style Augmentation (SA) Different from many methods whose style transfer networks leverage a style prediction network to generate style embeddings s, SA follows the style transfer network P which follows (Ghiasi et al. 2017) trained on the PBN dataset[1] and discards its style prediction network by sampling s from a multivariate normal distribution with the same mean μ and covariance Σ as the PBN datasets. In this way, the burden of style transfer networks can be significantly released. Simultaneously, to further constrain the strength of transformation, SA introduces the hyperparameter η_s to obtain the output embedding z:

$$z = \eta_s N(\mu, \Sigma) + (1 - \eta_s) X_T \tag{1}$$

Then, the embedding z which represents the new style, and x_T which is seen as a canvas will be incorporated again to generate feature maps. Therefore, the output feature maps M_s from SA can be expressed as:

$$M_s = \frac{\gamma(P(X_T, z) - \mu')}{\delta} + \beta \tag{2}$$

The mean and standard deviation across the feature map spatial axes are represented by μ' and δ, respectively, while γ and β represent the weight and bias obtained from the style transformer network.

Image Merging As mentioned earlier, to alleviate the decrease of content information caused by SA, we employ Image Merging to merge the original images with their augmented version, creating a combined input of our model. For the image merging of MSA, we introduce another random hyperparameter η_c which is randomly sampled from $U(0, 1)$ to merge the original image X_T and X'_T to build final training images X_M. We note that the usage of η_c is to decide the rate of merging position rather than simple multiplication.

$$X_M = Merge(\eta_c X_T, (1 - \eta_c) X_{T'}) \tag{3}$$

Severity Perception Mask (SPM)

To make further learning of consistent features from augmented scenes generated by MSA, we propose SPM to work on the source domain images and augmented scene images. SPM employs a Severity Perception mechanism and Mask Operation. The aim of Mask Operation is to further eliminate the weather-specific features of images from the source domain. However, a single Mask Operation will miss alignment of severity level leading to decreases in Cross-domain Consistency. Thus, we propose the Severity Perception mechanism to evaluate image severities and then strengthen the alignment of images with different severity.

Mask Operation As discussed before, we argue that the weather-specific features are related to four factors, including brightness, hue, noise, and contrast of images. Therefore, our masks consist of randomly sampling from uniform distributions of these factors. In this way, the factors of source domain images can be further randomized so that our model

[1]https://www.kaggle.com/c/painter-by-numbers

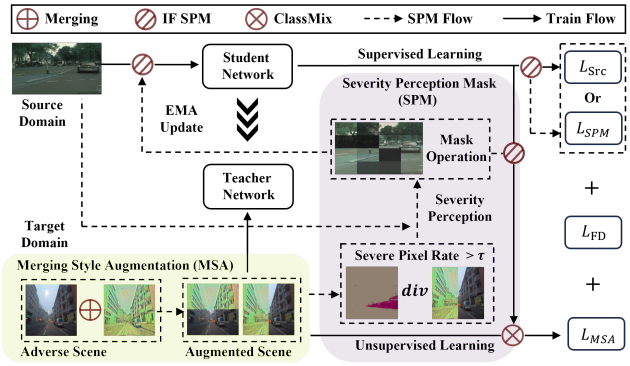

Figure 2: Training flow of PASS. First, target images will be augmented by MSA and then source domain images will be judged whether to be augmented by SPM decided on whether augmented target images pass the Severity Perception. And then student network will conduct supervised learning to generate $L_{Src/SPM}$. The teacher network will produce the pseudo labels of augmented target images to mix with the prediction of source images so that the student network can calculate unsupervised loss L_{MSA}.

can decrease the sensitivity to them. To achieve Mask Operation, we set min and max as the different minimum and maximum of every uniform distribution. Notably, our Mask Operation will divide an image into many blocks and the samplings of every block are also different. The first step for Mask Operation to generate masks is calculating the number N_b of blocks. We set the height and width of blocks are H' and W', and we denote the ideal number of blocks as N_i. Thus, the H' and W' can be expressed as:

$$N_b = \lfloor \sqrt{N_i} \rfloor \tag{4}$$

$$H'/W' = \frac{H/W}{N_b} \tag{5}$$

And then four factors will be randomly sampled from different uniform distributions and imposed on the original images X_S in sequence to generate masked image $X_{S'}$.

$$Mask = bright \bigotimes hue \bigotimes contrast \bigotimes noise \tag{6}$$

$$factors \sim U(min, max) \tag{7}$$

where $\bigotimes$ means the merging of different augmented factors and $factors$ including bright, hue, contrast, and noise. Thus the achievement of masked images is:

$$X_{S'}^{(H,W,3)} = \sum_{N_b} \sum_{h' \in H'} \sum_{w' \in W'} Mask(X_S^{(h',w',3)}) \tag{8}$$

Severity Perception Severity Perception mechanism unifies all augmented scene images X_M by measuring their severity by calculating their severity pixels ratio. To obtain the severity pixels ratio, we first transform the augmented images to their gray-scale map version $X_h \in \mathrm{R}^{\mathrm{H} \times \mathrm{W}}$, and then count the number of the gray pixels passing the severity value v to screen out severe pixels. Finally, we calculate

the ratio of severe pixels in all pixels and judge whether this ratio exceeds the severity threshold τ. If the ratio passes τ, the Mask operation will be activated on the X_S to generate $X_{S'}$ which is in the same training mini batch as X_M. The process of Severity Perception controlling Mask Operation is shown as follows:

$$Mask_{on} = \begin{cases} 1, & if \quad \frac{\sum_{h\in H}\sum_{w\in W}(X_h)^{(h,w)} > v}{H\cdot W} > \tau \\ 0, \end{cases} \tag{9}$$

where $Mask_{on}$ means the activation of Mask Operation.

Overall Training Flow

The whole training flow is shown in Fig 2. PASS consists of two networks which are a student network g_θ corresponding to the source domain and a teacher network g_ϕ corresponding to the target domain. Thus, the $X_{S/S'}$ is the input of g_θ and the image X_M is the input of g_ϕ. The main function of g_ϕ is to generate valuable pseudo labels $\hat{Y}_M$ of X_M and it does not participate in the training process. The key point of training is g_ϕ as it needs to perceive the Severity of input images to decide whether activate Mask Operation on X_S. g_ϕ will generate pseudo labels $\hat{Y}_M$ of augmented scene images X_M and respectively mix $X_{S/S'}$ with X_M and $\hat{Y}_{S/S'}$ with $\hat{Y}_M$ by ClassMix mechanism (Tranheden et al. 2021). g_θ is trained by calculating cross-entropy losses $L_{Src/SPM}$ and L_{MSA} and moves its parameters to the teacher network by EMA (Tarvainen and Valpola 2017) at the beginning of every iteration. Notably, the generation of $\hat{Y}_M$ follows our baseline (Hoyer, Dai, and Van Gool 2022b) which means that an estimated parameter q will be added to measure the confidence of $\hat{Y}_M$ due to the unsupervised learning of the teacher network.

$$q^c = \frac{\sum_{h\in H}\sum_{w\in W}(max_{c'} g_\phi(X_M^{(h,w,c')}) > \tau')}{H\cdot W} \tag{10}$$

where τ' is the constrained parameter and only when the predicted proportion of pixels in some categories of the target domain images exceeds the threshold τ', the prediction of these categories will become pseudo-labels. Thus, the L_{MSA} is:

$$L_{MSA} = -\sum_{H,W}\sum_{c\in C} q_M^c Y_M^{(H,W,c)} log\ g_\theta\left(X_M^{(H,W,c)}\right) \tag{11}$$

Although L_{MSA} needs q to constrain, there is no need to control the $\hat{Y}_S$ by introducing new hyperparameters because the student network conducts reliable supervised learning to obtain L_{SPM} calculated with masked images or L_{Src} calculated with original images.

$$L_{Src/SPM} = -\sum_{H,W}\sum_{c\in C} Y_{S/S'}^{(H,W,c)} log\ g_\theta\left(X_{S/S'}^{(H,W,c)}\right) \tag{12}$$

To obtain the final loss function, we follow the feature distance loss L_{FD} proposed in (Hoyer, Dai, and Van Gool 2022a) and combine all of the losses together:

$$L = L_{MSA} + L_{Src/SPM} + 0.005\cdot L_{FD} \tag{13}$$

Experiments

Datasets

Cityscapes (Cordts et al. 2016) Cityscapes (CS) contain 2,975 for training, 500 for validation, and 1,525 for testing of driving scenes captured in 50 urban areas. **ACDC (Sakaridis, Dai, and Van Gool 2021)** ACDC comprises four adverse scenes: fog, night, rain, and snow. For every scene, there are 400 images for training, 100 images for validation (including 106 night images), and 500 images for testing. **Foggy Zurich (Sakaridis et al. 2018)** Foggy Zurich (FZ) contains 1,522 images in the light fog and 1,498 images in the medium fog. **Foggy Driving (Sakaridis et al. 2018)** Foggy Driving (FD) comprises 101 real-world scenarios of foggy road conditions. It purely is a test benchmark. **Dark Zurich (Sakaridis, Dai, and Gool 2019)** Dark Zurich (DZ) comprises 8,779 images captured during nighttime, twilight, and daytime. It also includes 50 validation images and 151 test images. **Nighttime Driving (Dai and Van Gool 2018)** Nighttime Driving (ND) contains 50 nighttime images with coarsely annotated ground truth. **BDD100K-Night (Yu et al. 2020)** BDD100K-Night (BD) is a subset of the BDD100K segmentation dataset, consisting of 87 nighttime images with accurate segmentation labels.

Comparison Experiment Settings

Implementation details The default implementation of our methods is based on HRDA (Hoyer, Dai, and Van Gool 2022b), and follows the HRDA-based implementation of the teacher-student self-training framework of DAFormer (Hoyer, Dai, and Van Gool 2022a), which includes feature distance loss, confidence-weighted pseudo labels ($\tau' = 0.968$), rare class sampling, and ClassMix following DACS (Tranheden et al. 2021). The optimizer used is AdamW (Loshchilov and Hutter 2017), with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, and a linear learning rate warm-up. Regarding the resolution setup details, we follow the default configuration and parameters of HRDA. Unless otherwise stated, the MSA parameters are set to $\eta_s = 0.5$ and $\eta_c = 0.5$, which means that X_T will be combined with $X_{T'}$ at a medium position as shown in Fig 2. For the Mask Operation of SPM, we set the min and max: $(-1.6, 1.6)$ for brightness, $(-15, 15)$ for hue, $(0.8, 1.2)$ for contrast, and $(0, 50)$ for noise. we also set $N_i = 24$, and $v = 40$ for the Severity Perception.

Fog-specialized Models Comparison To demonstrate our SOTA performance under fog scenes, we conduct CS to ACDC-Fog, CS to FZ, and generalize to FD which means the source domain uses CS, and the target domain uses ACDC or FZ datasets. Notably, we train PASS on FZ and test on FZ and FD. We compare PASS with fog-specialized models, including SFSU (Sakaridis, Dai, and Van Gool 2018), CMAda3 (Dai et al. 2020), CuDA-Net (Ma et al. 2022), FIFO (Lee, Son, and Kwak 2022), FogAdapt (Iqbal, Hafiz, and Ali 2022). We set the $\tau = 0.08$ in CS to FZ and $\tau = 0.05$ in CS to ACDC-Fog. The results are shown in the foggy scenes of Table 1.

Night-specialized Models Comparison To show our performance under night scenes, we conduct experiments on

Models	Adverse Scenes (mIoU)									
	Foggy			Night				Rainy	Snowy	All
	ACDC-Fog	FZ	FD	ACDC-Night	DZ	ND	BD	ACDC-Rain	ACDC-Snow	ACDC-All
Specialized Models										
SFSU	-	35.7	46.3	-	-	-	-	-	-	-
CMAda3	-	46.8	49.8	-	-	-	-	-	-	-
CuDA-Net	55.6	49.1	*53.5*	-	-	-	-	-	-	-
FIFO	-	48.4	50.7	-	-	-	-	-	-	-
FogAdapt	-	*50.6*	53.4	-	-	-	-	-	-	-
GCMA	-	-	-	-	42.0	45.6	33.2	-	-	-
MCGDA	-	-	-	-	42.5	49.4	34.9	-	-	-
DANNet	-	-	-	-	44.3	47.7	-	-	-	50.0
Semi-Unified Models										
AdaptSeg	-	26.1	37.6	-	30.4	34.5	22.0	-	-	-
DAFormer	48.9	40.8	-	44.7	48.5	51.8	33.9	59.9	53.7	55.4
SePiCo	58.5	-	-	50.5	54.2	*56.9*	*40.6*	66.1	57.9	59.1
HRDA	*69.9*	46.0	-	*53.1*	55.9	-	-	*73.6*	*69.5*	68.0
MIC	-	49.7	-	-	*60.2*	-	-	-	-	*70.4*
Unified Models										
PASS (Ours)	**70.6**	**59.9**	**60.2**	**60.3**	**60.2**	**57.0**	**43.0**	**74.6**	**70.0**	**70.8**

Table 1: Model-level comparison with PASS and other UDA for semantic segmentation methods. We compare PASS with the state-of-the-art specialized models for foggy scenes and night scenes, and Semi-Unified models which test on a part of adverse scene benchmarks. **Bold** denotes the best result and *italics* denotes the second-best.

CS to ACDC-Night, CS to DZ, and generalize to ND and BD. We compare PASS with night-specialized models, including GCMA (Sakaridis, Dai, and Gool 2019), MCGDA (Sakaridis, Dai, and Van Gool 2020), and DANNet (Wu et al. 2021). We set $\tau = 0.03$ in CS to DZ, $\tau = 0.035$ in BN, $\tau = 0.055$ in ND, and $\tau = 0.05$ in CS to ACDC-Fog. The comparisons are shown in the night scenes of Table 1.

Semi-Unified Models Comparison We define Semi-Unified models as representing the previous models that conduct tests on the part of benchmarks under different scenes. We compare PASS with these Semi-Unified models including AdaptSeg (Tsai et al. 2018), DAFormer (Hoyer, Dai, and Van Gool 2022a), SePiCo (Xie et al. 2023), HRDA (Hoyer, Dai, and Van Gool 2022b), and MIC. Except for the foggy and night scenes experiments, we conduct CS to ACDC-Rain, CS to ACDC-Snow, and CS to ACDC-All Conditions. In these experiments, we also set the threshold parameter $\tau = 0.05$.

Performance Experiments Analysis

We present the state-of-the-art performance of PASS in Table 1. The results in foggy scenes demonstrate that our PASS surpasses previous fog-specialized models and Semi-Unified models across three foggy benchmarks. In the Cityscapes to ACDC-Fog, PASS outperforms the Semi-Unified model HRDA by 0.7% mIou. Particularly, for Cityscapes to FZ and generalizing to FD, our PASS achieves a 9.3% mIou improvement over FogAdapt in FZ and a 6.7% mIou improvement over CuDA-Net in FD. These findings indicate that PASS exhibits excellent adaptation capabilities to foggy scenes and even outperforms specialized models. For further analyses, we visualize the predictions under various adverse scenes in Fig 6. From the visualizations, we observe that our PASS excels in predicting the sky, which is often misclassified as vegetation by HRDA, leading to a significant improvement in FZ.

Focusing on the comparison under night scenes, PASS significantly outperforms night-specialized models and achieves comparable performance to the Semi-Unified model MIC in the Cityscapes to DZ. Moreover, in the generalization to ND and BD, PASS demonstrates advantages over SePiCo with a 0.1% improvement in ND and a 2.4% improvement in BD. Notably, in the Cityscapes to ACDC-Night, PASS surpasses our baseline model HRDA by 7.3% mIou. Night scene tasks are typically the most challenging in UDA for semantic segmentation, but our PASS still exhibits strong robustness. We attribute this success to SPM, which enhances the ability of PASS to tolerate low-light conditions and adapt to night scenes. These results align well with the observation that the content information in night scenes is the most prominent among all adverse scenes, as shown in Fig 3 (b). Visualizing the predictions under ACDC night in Fig 6, we can see that compared to the baseline, our PASS learns more consistent features.

Regarding the performance in rainy and snowy scenes, since there are no specialized models dedicated to these two scenes, we compare PASS with semi-unified models. The results reveal that PASS achieves state-of-the-art performance by surpassing HRDA by 0.1% mIou in ACDC-Rain and 0.5% mIou in ACDC-Snow. The last two predictions in Fig 6 correspond to rainy and snowy scenes, respectively, and they demonstrate that PASS effectively learns consistent features, such as tiny traffic lights, which are more challenging to learn compared to other classes. Finally, in the Cityscape to ACDC-all conditions experiments, our PASS also achieves state-of-the-art results.

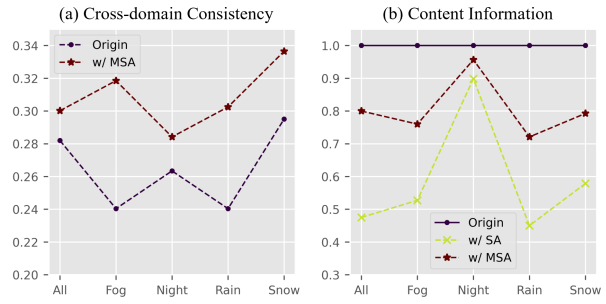

Figure 3: (a) means that MSA can significantly increase the Cross-domain Consistency by eliminating the weather-specific features. (b) quantifies the content information which is reduced by SA and strengthened by MSA.

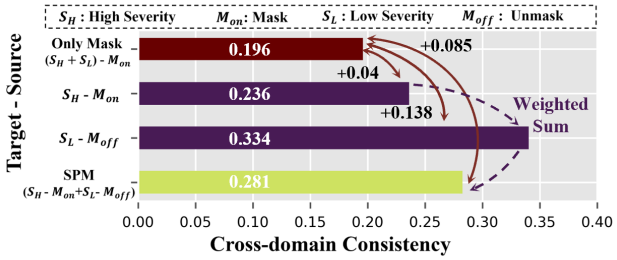

Figure 4: We validate that Severity Perception improves Cross-domain Consistency. Only Mask indicates low consistency caused by the single Mask Operation. S_H-M_{on} shows that images with high severity exhibit higher consistency with masked images. S_L-M_{off} demonstrates that images with low severity also exhibit increased consistency with unmasked images. SPM demonstrates that Severity Perception effectively improves consistency.

Cross-domain Consistency Analysis

As we discussed above, unlike ambiguous descriptions of consistency in previous works, we will quantify our Cross-domain Consistency by calculating SSIM. Thus, we hope to illustrate how the influence of Cross-domain Consistency brought by MSA and SPM. We note that all Cross-domain Consistency is calculated by the average SSIM of hundreds of pairwise images.

Firstly, we demonstrate the influence of Cross-domain Consistency brought about by MSA. In Fig 3 (a), we calculate consistency between the Cityscapes source domain and four ACDC target domains. The abscissa represents the four adverse scenes of ACDC under various conditions. It is evident that the utilization of MSA to augment images significantly enhances Cross-domain Consistency compared with the original images. Since we formulated that SA may result in a decrease in content information, we incorporate the Image Merging mechanism to address this issue. Thus, we further analyze the content information by separating the structure from the SSIM and calculating the structure under

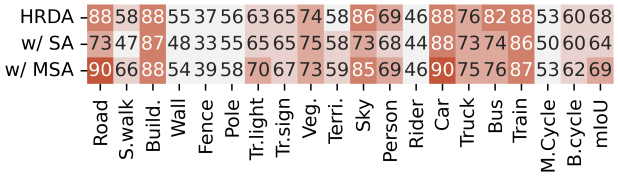

Figure 5: Ablation studies of SA and MSA on Cityscapes to ACDC by heat map visualization. The results show that content information brought by MSA significantly improves performance.

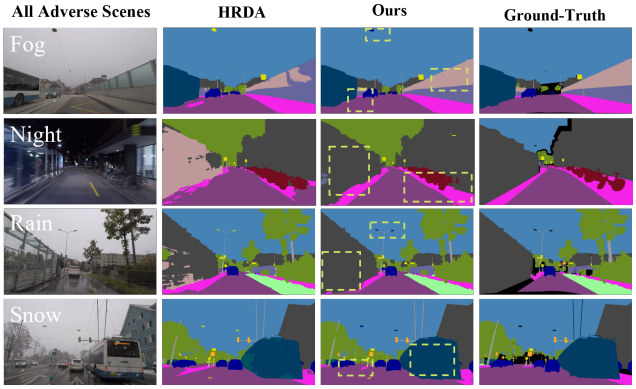

Figure 6: Visualization of prediction of PASS and baseline under all adverse scenes. Our PASS is significantly superior to the baseline.

each scene. As illustrated in Fig 3 (b), SA leads to a noticeable decrease in content information across all scenes. After incorporating the Image Merging mechanism to build MSA, the decreasing trend is significantly alleviated across all scenes. Notably, in the night scene, the content information reverts back to 0.956 from 0.898.

Based on the enhancement of Cross-domain Consistency from MSA, we also hope our SPM can also ensure a high-level Cross-domain Consistency. We argue that employing a single Mask Operation can strengthen the sensibility of our model to weather-specific features on the source domain, but it will cause an imbalance of the corresponding relationship between the images with different severity in the training mini-batch, and further reduce the Cross-domain Consistency. By introducing Severity Perception, we strengthen the corresponding relationships again to enhance Cross-domain Consistency by aligning the images with similar severities. We show the quantified results in Fig 4.

In Fig 4, the abscissa values Cross-domain Consistency, while the ordinate means that Cross-domain Consistency will be calculated between augmented scene images (target domain) and masked or unmasked images (source domain). Only Mask means the consistency between all target domain images (S_H + S_L) with the masked images M_{on} indicating that using a single Mask Operation will cause the imbalance leading to a decrease in Cross-domain Consistency. S_H-M_{on} represents the consistency between the se-

	Bri.	Hue	Noi.	Contr.	mIou	gain
1	-	-	-	-	68.0	-
2	✓	-	-	-	68.8	+0.8
3	✓	✓	-	-	68.8	+0.8
4	✓	✓	-	✓	69.9	+1.9
5	✓	✓	✓	-	70.1	+2.1
6	✓	✓	✓	✓	70.8	+2.8

Table 2: Ablation studies on the four factors of Mask Operation to validate the effectiveness. These experiments are based on using MSA and Severity Perception. And the results demonstrate the combination of four factors can achieve the best performance.

	MSA	Mask Op.	Seve.	mIou	gain
1	-	-	-	68.0	-
2	✓	-	-	68.7	+0.7
3	-	✓	-	62.2	-5.8
4	-	✓	✓	67.5	-0.5
5	✓	✓	-	68.2	+0.2
6	✓	✓	✓	70.8	+2.8

Table 3: Ablation studies on MSA and mechanisms of SPM. The results show the necessity of MSA, Mask Operation, and Severity Perception.

vere images and masked images is higher than Only Mask by 0.04, indicating that Severe images align better with masked images than all target images. Furthermore, S_L-M_{off} also demonstrates that low-severity images align better with unmasked images which also possess low severity. Finally, SPM, the weighted sum of S_H-M_{on} and S_L-M_{off}, effectively demonstrates that Severity Perception aids in Mask Operation achieving 0.085 improvements of Cross-domain Consistency compared to Only Mask.

Therefore, we reasonably conclude that both MSA and SPM significantly enhance Cross-domain Consistency by their different functions. MSA enhances consistency by effectively eliminating weather-specific features and SPM enhances consistency by employing Severity to guide our model to align images with different severities well.

Ablation Study Analysis

We conduct ablation studies comparing the original SA with our MSA, which includes SA and Image Merging, as illustrated in Fig 5. The results clearly demonstrate that MSA outperforms SA by 5% mIou and surpasses HRDA by 1%. Notably, SA performs even worse than HRDA, with a 4% mIou decrease. This observation aligns with the negative impact of SA on image content information, which HRDA is particularly sensitive to. In contrast, MSA effectively addresses this issue. Additionally, by analyzing the improvements in classes such as Traffic Light, Fence, and Bicycle, we can infer that enhanced Cross-domain Consistency significantly assists our model in accurately recognizing challenging-to-learn classes.

We further conduct ablation studies on MSA and mechanisms of SPM on Cityscapes to ACDC-All Conditions, as presented in Table 2. The baseline performance is represented in the first row for comparison purposes. From the second row, we observe that employing a single MSA results in a 0.7% gain by enhancing Cross-domain Consistency from the target domains. Next, we focus on the third row, which indicates that using a single Mask Operation without MSA leads to a significant decrease of -5.8% in performance. We argue that this decrease is primarily due to the lack of consistency. However, when we incorporate Severity Perception to strengthen Cross-domain Consistency by balancing the severity level, this decrease is mitigated from -5.8% to -0.5%. Furthermore, we can observe the added value contributed by consistency in the fifth and sixth rows, where MSA aids both the single Mask Operation and SPM, resulting in performance improvements of -5.8% to 0.2% and -0.2% to 2.8%, respectively. The combining of MSA and SPM also outperforms the using single MSA by 2.1%. Therefore, both MSA and SPM are essential components in constructing PASS.

As we discussed above, our Mask Operation consists of randomly sampling four emphases, brightness, hue, noise, and contrast to decrease the sensibilities of our model to weather as weather-specific features are always related to these four emphases. Therefore, we are going to conduct ablation studies on Cityscapes to ACDC-All Conditions to validate the effectiveness of constitutions of Mask Operation. In Table 3, we still set the performance of HRDA in the first row as the base comparison. Notably, the usage of the MSA and Severity Perception are default conditions. From the second row, we observed a 0.8% gain when using single brightness sampling. However, the third row showed no improvement when combining brightness with hue. To further investigate the effectiveness of hue, we conducted the fourth and fifth rows. The results demonstrated that the performance in the fifth row, which included hue, was better than in the fourth row without hue. Additionally, the sixth and fifth rows confirmed the effectiveness of contrast. Based on these findings, we can conclude that hue needs to work in conjunction with noise or contrast, and that each emphasis is necessary to achieve unification. Experiments on parameter sensitivity to τ and N_i are shown in the appendix.

Conclusion

In this paper, we aim to create a unified model named PASS to parse all adverse scene tasks in UDA for semantic segmentation. To achieve this, we propose a new measurement, Severity, to lead our model to focus on the learning of highly consistent features rather than weather-specific features of all adverse scenes. In this process, we conduct careful analysis experiments and ablation studies on the quantification of our Cross-domain Consistency to validate why our method can work well and how we create MSA and SPM. We also conduct sufficient performance experiments on almost widely used benchmarks including ACDC (fog, night, rain, and snow), Foggy Zurich, Foggy Driving, Dark Zurich, Nighttime Driving, and BDD100K-Night.

Acknowledgements

This work was supported by the Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System.

References

Brüggemann, D.; Sakaridis, C.; Truong, P.; and Van Gool, L. 2023. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3174–3184.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Dai, D.; Sakaridis, C.; Hecker, S.; and Van Gool, L. 2020. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128: 1182–1204.

Dai, D.; and Van Gool, L. 2018. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3819–3824. IEEE.

Gao, H.; Guo, J.; Wang, G.; and Zhang, Q. 2022. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9913–9923.

Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; and Shlens, J. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*.

Gong, Z.; Li, F.; Deng, Y.; Shen, W.; Ma, X.; Ji, Z.; and Xia, N. 2023. Train One, Generalize to All: Generalizable Semantic Segmentation from Single-Scene to All Adverse Scenes. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2275–2284.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, 1989–1998. Pmlr.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022a. 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*, 9924–9935.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022b. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, 372–391. Springer.

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11721–11732.

Iqbal, J.; Hafiz, R.; and Ali, M. 2022. FogAdapt: Self-supervised domain adaptation for semantic segmentation of foggy images. *Neurocomputing*, 501: 844–856.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Lee, S.; Son, T.; and Kwak, S. 2022. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18911–18921.

Lei, Y.; Emaru, T.; Ravankar, A. A.; Kobayashi, Y.; and Wang, S. 2020. Semantic image segmentation on snow driving scenarios. In *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, 1094–1100. IEEE.

Li, R.; Tan, R. T.; and Cheong, L.-F. 2020. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3175–3185.

Liao, Y.; Zhou, W.; Yan, X.; Cui, S.; Yu, Y.; and Li, Z. 2022. Geometry-Aware Network for Domain Adaptive Semantic Segmentation. *arXiv preprint arXiv:2212.00920*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, X.; Wang, Z.; Zhan, Y.; Zheng, Y.; Wang, Z.; Dai, D.; and Lin, C.-W. 2022. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18922–18931.

Sakaridis, C.; Dai, D.; and Gool, L. V. 2019. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7374–7383.

Sakaridis, C.; Dai, D.; Hecker, S.; and Van Gool, L. 2018. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, 687–704.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2020. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3139–3153.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10765–10775.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tranheden, W.; Olsson, V.; Pinto, J.; and Svensson, L. 2021. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1379–1389.

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

Valanarasu, J. M. J.; Yasarla, R.; and Patel, V. M. 2022. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2353–2363.

Wang, X.; Wu, Y.; Zhu, L.; and Yang, Y. 2020. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12249–12256.

Wang, X.; Zhu, L.; and Yang, Y. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5079–5088.

Wang, X.; Zhu, L.; Zheng, Z.; Xu, M.; and Yang, Y. 2022. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, X.; Wu, Z.; Guo, H.; Ju, L.; and Wang, S. 2021. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15769–15778.

Xie, B.; Li, S.; Li, M.; Liu, C. H.; Huang, G.; and Wang, G. 2023. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.