Supplementary Material for
IDD-AW : A Benchmark for Safe and Robust Segmentation of Drive Scenes in Unstructured Traffic and Adverse Weather

1. Experimental Details

1.1. Dataset

We conduct experiments on various road scene datasets ACDC, IDD and our IDD-AW dataset. IDD contains images from unstructured Indian road conditions whereas ACDC dataset contains images from 4 weather conditions Rain, Fog, Night time and Snow. Our dataset IDD-AW contains images from 4 adverse conditions Rain, Fog, Lowlight and Snow as well as contains images in various unstructured road scenes.

1.2. Semantic Segmentation

Training Details

The first task IDD-AW supports is standard semantic segmentation. For inference and training of each of the models, we use InternImage framework, which aligns with the current leading standard for Cityscapes dataset. In this task, we train semantic segmentation using various combinations of train and test sets to compare the results among them. We have state-of-the-art pretrained models on the above mentioned datasets and these are evaluated on the IDD-AW test set. All model variants are trained on an input size of 1024x768 for 160K/320k iterations. For the optimizer, we use Adam optimizer with a momentum of 0.9 and an initial learning rate of 0.0001. We also use Cross Entropy Loss as the loss function.

Evaluation of Pre-trained Models on IDD-AW

For evaluation, we have taken state-of-the-art network of InternImage framework trained on IDD Dataset, ACDC and IDD-AW dataset. For each of the existing datasets, we have taken the most recent model, InternImage with different backbones and have performed inference on our test set. The input image size for each of the experiment is 1024x768. For quantitative comparision of various models and datasets, refer to table 2 in the paper. Additionally, the tables 1 and 2 are experiments using InternImsge-s. The similar experiment results put in the main paper are using InternImage-b.

Dataset Test → Train ↓	CS	ACDC	IDD	Rain	Fog	LL	Snow	IDD-AW
CS RGB	-	-	-	43	48	42	39	45
ACDC RGB	-	-	-	45	49	40	37	45
IDD RGB	-	-	-	51	54	49	33	52
IDD-AW RGB	-	-	-	62	64	62	49	63
IDD-AW NIR	-	-	-	60	56	56	50	60
IDD-AW NIR+RGB	-	-	-	**65**	**64**	**62**	**53**	**66**

Table 1. Comparison of mIoU scores (%) of InternImage-s model trained on CityScapes (CS), ACDC, IDD, and IDD-AW datasets (RGB, NIR, and Combined) for IDD-AW test set in individual conditions and jointly for all conditions. The model trained on IDD-AW NIR+RGB gives the highest accuracy which is 20% more than CS, ACDC, and 14% more than IDD. The NIR+RGB model gives 3% more accuracy compared to the RGB model, which indicates that the NIR image adds useful information for prediction.

	Condition	mIoU	SmIoU (tp)	SmIoU
InternImage-s	All	63.42	57.55	47.26
	Rain	61.22	55.71	45.11
	Fog	64.02	57.50	46.7
	Lowlight	61.70	57.35	46.27
	Snow	49.22	38.51	23.33

Table 2. Comparison of mIoU(%) with SmIoU (%) metric at different levels and label sets for various adverse weather conditions. Here, SmIoU refers to Safe mIoU, and tp refers to just the traffic participants

Qualitative Results

From Fig 1, we compare various models trained on ACDC, IDD and IDD-AW dataset. Here, we focus on the sidewalk class and how its dangerously mispredicted in each of the

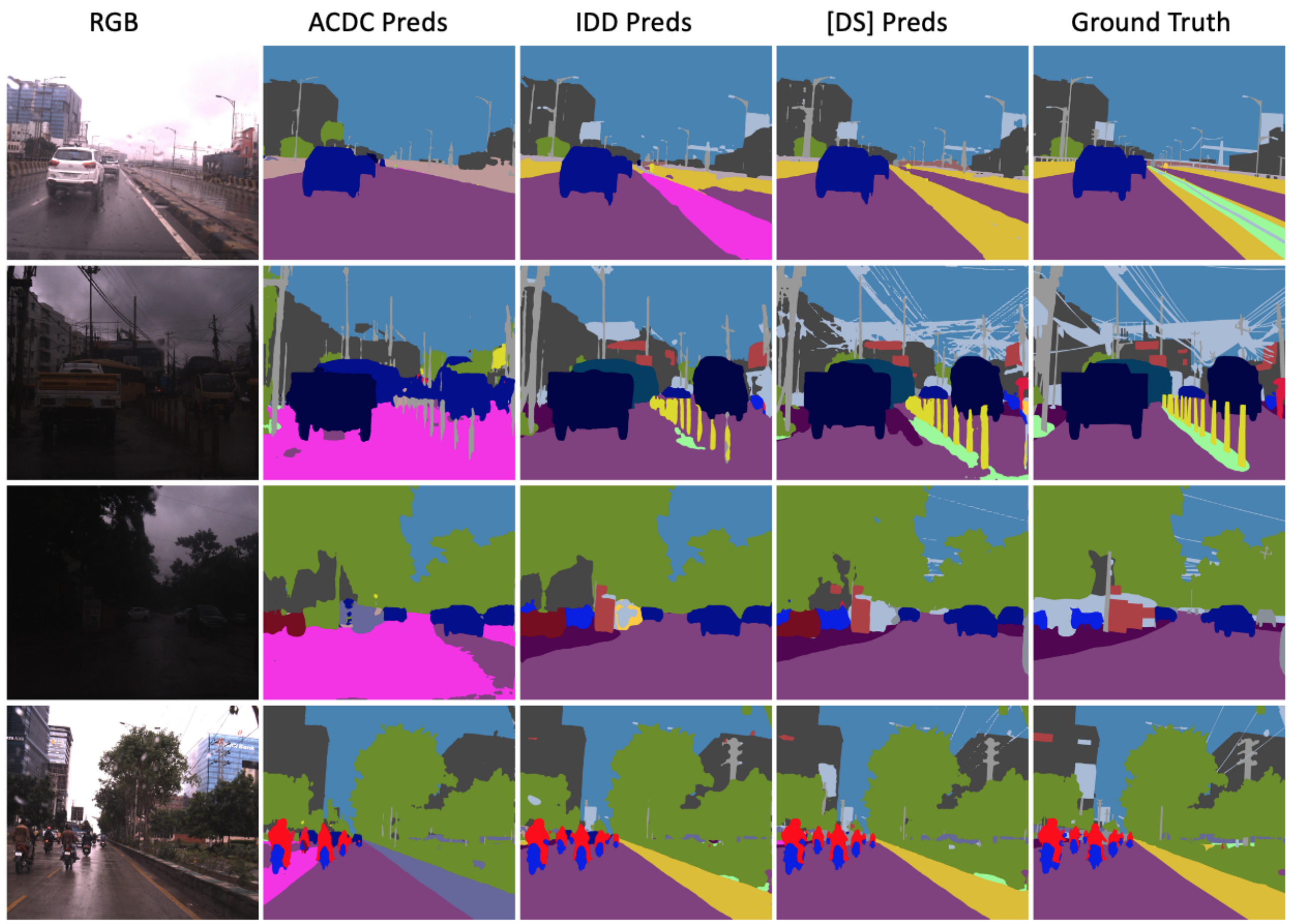

Figure 1. This is a predictions comparision between ACDC vs IDD vs IDD-AW pretrained model tested on IDD-AW test set. The above image shows the importance of sidewalk class. The ACDC pretrained model cannot identify sidewalk in the first image, whereas it identifies the whole road or part of the road as the sidewalk itself. This is a dangerous misclassification as it can lead to disastrous results when deployed as realtime.

images. In the first image, the sidewalk is not at all predicted in the ACDC model, whereas our pretrained model clearly predicts the sidewalk as evident in the ground truth. For the next couple of images, the road itself is misclassified as sidewalk partially or completely, which is unacceptable, especially because road is a drivable surface whereas sidewalk is a non-drivable. Here, even though there are vehicles clearly visible on the road, the ACDC model mispredicts it as sidewalk. This is one of the many examples to show the superiority of our new dataset when compared with other adverse weather datasets.

2. Safe mIoU in more detail.

To know the Safe mIoU, first we need to know about the tree distance. The tree distance is a measure of distance between two classes in a hierarchical dataset. When both the classes have the same immediate parent, the tree distance between the two classes becomes 1. Similarly, if two classes do not share the same parent, but share the same grandparent node in the hierarchical tree, the tree distance between the two classes is 2. And it goes on based on the number of levels of the dataset.

We explain the above using the pictorial representation of our dataset. The IDD-AW dataset follows the same label set as IDD. Hence, it follows four levels of hierarchy. Now with reference to figure 5 in the main paper, truck, bus and vehicle fallback share the same parent. Hence, the tree distance between those classes = 1. Here, the classes with tree distance 1 is notated with a yellow egdes to show the least level of severity while being misclassified. We can see this from the fig 2 in fog image, where the truck and vehicle fallback are misclassified as one another, hence the severity map shows yellow color.

Similarly, rider and person do not share the immediate parent, but come under the same L1 label, Living things.

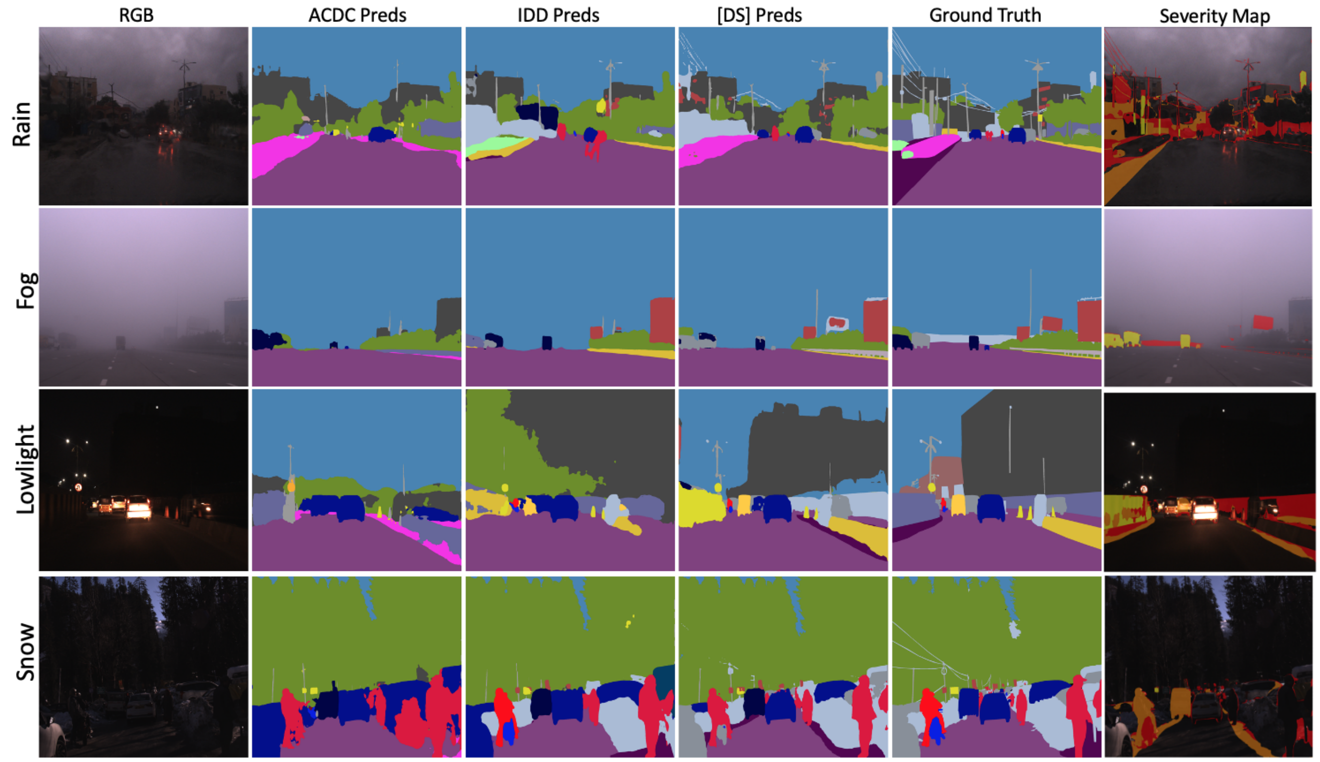

Figure 2. The image illustrates predictions made on the IDD-AW test set using pretrained models from ACDC, IDD, and IDD-AW , along with ground truth and severity maps generated by the IDD-AW pretrained model's predictions. In the severity maps, colors signify various danger levels, where yellow indicates misclassification at Level 3 (the lowest level of the tree), orange represents misclassification at Level 2, and red corresponds to Level 1, collectively indicating the overall danger level of the driving scene.

Hence, the tree distance between those classes is 2. This is shown by the orange dotted edges between Person and rider. Same as sidewalk and non drivable fallback which have a tree distance of 2. We can see this misclassification from fig 2, where the rain image has misclassification between sidewalk and non drivable fallback and hence the severity map shows orange color. Similarly, for the snow image, rider is misclassified as person and car is misclassified as truck. Hence, the severity map shows those labels in orange color.

For the farthest classes which do not have any common nodes in any level are represented by the red color edges.

2.1. Calculation of Safe mIoU

To calculate Safe mIoU, we first define a set of important classes in our dataset.We calculate our traditional mIoU on these complete 26 classes label set. However, we calculate the safe metric using the metric in eq 2 from the paper, defined previously.

2.2. Quantitative results

As depicted in Figure 1 (Figure 2), a qualitative analysis reveals a marked superiority in the performance of the IDD-AW model compared to the ACDC and IDD models. Detailed quantitative results for the ACDC, IDD, and IDD-AW datasets are presented in Table 2 of the main paper.

Specifically, when utilizing the IDD-AW pre-trained model for the images in Figure 1, the mean Intersection over Union (mIoU) for rain is measured at 39.03. However, it's worth noting that the Safe mIoU (SmIoU) registers at 8.15. This significant difference can be attributed to misclassifications at various levels: car, bike, and rider at Level 1, and sidewalk at Level 2, all of which constitute hazardous conditions for driving in inclement weather.In foggy conditions, misclassifications include billboard at Level 3, car at Level 2, and truck and vehicle fallback at Level 1. For images with the IDD-AW pre-trained model, the mIoU and SmIoU for fog are 52.92 and 40.76, respectively. In low light conditions, misclassifications involve a wall at Levels 2 and 3, partial misclassification of curb at Levels 1 and 2, and a section of a car at Level 2. The mIoU and SmIoU for these images are 36.18 and 26.43, respectively. Lastly, in snowy conditions, misclassifications encompass the merging of riders with persons, misclassification of distant persons and far objects at Level 3, and misclassification of roadsides and cars at Level 2. The mIoU and SmIoU for these images are 50.09 and 31.95, respectively.

These findings highlight the significance of SmIoU in quantifying safety aspects within driving scenarios. SmIoU provides a more comprehensive evaluation, assigning higher penalties to orange and red misclassifications, thereby aligning more closely with safety considerations within driving scenes compared to traditional mIoU.