# ERF: A Benchmark Dataset for Robust Semantic Segmentation Under Extreme Rainfall Conditions

**Xin Yang**[1], **Xin Zhang**[1], **Xinchao Wang**[1]

[1]National University of Singapore
e0674612@u.nus.edu, x.zhang@u.nus.edu, xinchao@nus.edu.sg

## Abstract

As climate change reshapes global weather patterns, the increasing frequency and intensity of extreme rainfall events have amplified the safety imperatives for autonomous driving systems. During such events, rainfall can escalate from heavy to violent, as defined by the World Meteorological Organization, severely impairing images with diverse and significant degradations. Many existing semantic segmentation models perform well under light to heavy rain, but there is a notable absence of datasets addressing violent rain conditions for these models to validate and learn from. In this paper, we introduce the Extreme RainFall (ERF) dataset for semantic segmentation in both image and video tasks under violent rain conditions. Our dataset comprises 14,757 unlabeled frames and 100 labeled frames, all captured during four different violent rainfall periods. We use our dataset to evaluate the robustness of various methods against violent rainfall, focusing on four approaches: 1) image-based foundation models, 2) image-based domain generalization methods, 3) image-based domain adaptation methods, and 4) video-based methods. The results reveal that none of the existing models tested is capable of withstanding the extreme challenges posed by violent rainfall conditions. By analyzing the results, we offer insights and suggestions for developing more robust models under extreme rainfall events.

## Introduction

In an era where climate change is reshaping weather patterns globally, the escalating frequency and intensity of extreme weather events have amplified the safety imperatives for autonomous driving systems. Consequently, developing robust vision models that can withstand extreme weather conditions has become a critical priority.

The World Meteorological Organization (WMO) categorizes rainfall intensity based on hourly precipitation rates into light rain ($< 2.5\mathrm{mm/h}$), moderate rain ($2.5 - 7.6\mathrm{mm/h}$), heavy rain ($7.6 - 50\mathrm{mm/h}$), and violent rain ($> 50\mathrm{mm/h}$). Under different rainfall conditions, visual degradation can vary significantly, as illustrated in Fig. 1. Therefore, having datasets that capture various rain conditions is essential to ensure the robustness of computer vision models. However, existing rainfall datasets for semantic segmentation tasks, whether synthesized or real-world, such as ACDC (Sakaridis, Dai, and Van Gool 2021) and BDD100K (Yu et al. 2020), predominantly capture light rain scenarios (Tung et al. 2017; Caesar et al. 2020; Yu et al. 2020; Sakaridis, Dai, and Van Gool 2021; Jin et al. 2021; Ji et al. 2023). Currently, only one real-world dataset (Zhong et al. 2022) includes images of heavy rain, and none provide data for violent rain conditions.

As climate change progresses, violent rain is becoming increasingly prevalent in regions historically unaccustomed to such extreme precipitation. For instance, during the 7.20 extreme rainfall event in Zhengzhou, China, hourly precipitation exceeded 201.9 mm, over four times the threshold for heavy rain. Under violent rainfall conditions, visual degradation can significantly increase, as illustrated in Fig. 1. Such degradation distorts and impairs visual appearances, posing a significant challenge to existing computer vision models. Despite the demonstrated efficacy of current computer vision models on existing datasets, their performance under violent rain conditions remains untested and uncertain. This gap poses substantial risks: Can our current transfer learning methods (e.g., domain generalization, domain adaptation) be trusted to function under violent rain conditions? Will widely adopted foundation models sustain their effectiveness in a world where violent rain events become more frequent?

Neglecting this risk is like constructing a dam without accounting for potential flood levels—an oversight with potentially catastrophic consequences. This issue is not just a future concern. Nearly 3 billion people reside in tropical regions where violent rain is already common. The pressing question is whether current computer vision models can reliably operate in these high-risk areas.

To address this critical gap, we follow the example of prior datasets for semantic segmentation tasks and propose the first real-world urban scene dataset captured under violent rainfall conditions, tailored for both image and video tasks. By rigorously evaluating existing models against this dataset, we aim to uncover their performance under extreme weather conditions, thereby enhancing the safety and reliability of autonomous driving systems in an increasingly unpredictable climate. Our dataset contains 14,757 unlabeled frames and 100 labeled frames, captured under four different violent rainfall conditions. In the ERF dataset section, we detail the meticulous collection and selection process to en-

(a) Light (ACDC) (b) Heavy (Rainy WCity) (c) **Violent (Ours)** (d) **Violent (Ours)**

Figure 1: Examples of images under different rainfall intensities reveal that the types and severity of degradations vary significantly. It is evident that violent rainfall conditions result in the most impaired visual appearances.

sure the dataset represents violent rain conditions. We also provide a comparison between our dataset and existing related datasets to highlight its unique attributes. In the Experiments section, we use our dataset to evaluate the performance of various common models and methods, categorized as follows: 1) Image-based foundation models, which are large-scale models trained on extensive datasets and designed for versatility across different domains. 2) Image-based domain generalization methods, which are trained under ideal conditions and aim to remain robust in adverse conditions. 3) Image-based domain adaptation methods, which are trained using images from both ideal and adverse conditions to achieve robustness in both scenarios. 4) Video-based semantic segmentation methods, which incorporate temporal information to generate accurate segmentation maps for the current frame. We summarize the contributions of this work as follow:

- We propose the Extreme RainFall (ERF) dataset, the first dataset under violent rainfall conditions, specifically designed for both image and video semantic segmentation tasks.
- We benchmark various models and methods across four different categories, evaluating their robustness against our ERF dataset during extreme rainfall events. Based on these evaluations, we offer insights and suggestions for potential directions to develop models that are more robust under such challenging adverse conditions.

## Related Work

The Cityscapes dataset (Cordts et al. 2016) is one of the most widely used datasets for urban scene understanding. It provides annotations for 5,000 images captured in 50 cities during daytime under ideal conditions. While it is comprehensive for various urban scenes, it lacks adverse weather conditions, limiting its utility for training models to handle challenging environments like rain (Li, Kou, and Zhao 2021; Özdenizci and Legenstein 2023; Chen et al. 2025).

RainCityscapes (Hu et al. 2019) extends the Cityscapes dataset by introducing synthetic rain effects to the original images. While it offers a valuable benchmark for rain-affected urban scenes, the synthetic nature of the rain may not fully capture the complexity of real-world rain. The Raincouver dataset (Tung et al. 2017) includes images captured in rainy conditions. This dataset provides annotations for object detection and segmentation, focusing on how rain affects visibility and the appearance of objects in urban scenes. The nuScenes dataset (Caesar et al. 2020), BDD100K dataset (Yu et al. 2020), and ACDC dataset (Sakaridis, Dai, and Van Gool 2021) are comprehensive image datasets for autonomous driving, including various adverse conditions such as rain, fog, and snow. The RaidaR dataset (Jin et al. 2021) is specifically designed for rain-affected driving scenes, providing the largest amount of real-world images captured in various rainy conditions. The MVSS dataset (Ji et al. 2023) is a video dataset that provides different urban scenes under various conditions, including rain. The Rainy WCity dataset (Zhong et al. 2022) offers a collection of urban scenes captured under heavy rainy conditions. While these existing datasets provide valuable resources for training and evaluating models under various rainfall conditions, there is a gap in datasets that specifically address the challenges posed by violent rainfall. Our dataset aims to fill this gap by offering comprehensive annotations for urban scenes captured under extreme rain events, providing a crucial resource for developing robust vision models capable of handling violent rainfall conditions.

## ERF Dataset

**Data Collection and Selection** Our goal is to collect an urban scene dataset under violent rainfall conditions. Therefore, we chose Singapore, a tropical country where violent rain is frequent. According to the latest World Bank data, Singapore ranks as the 14th country globally in terms of average precipitation depth (Bank 2024). Additionally, the country has a network of over 60 weather stations that record rainfall intensity across various regions (Singapore 2024). By checking the recorded intensity from the nearest station to our data collection area, we can accurately identify the

corresponding rainfall level. While the primary reason for choosing Singapore is its frequent violent rain, it is also a developed nation with the potential for widespread adoption of autonomous driving and diverse traffic scenarios, further enhancing the relevance of our dataset.

To capture the necessary data, we actively respond to rainfall events. When a significant rainfall event begins in Singapore, we drive to the affected region to collect video footage. We mount an OAK-D camera inside the car, positioned on the roof and behind the windshield. This camera system includes a main color camera and a pair of stereo lenses. The main color camera captures video at 15 frames per second, while the stereo lenses simultaneously collect video from different views, generating a corresponding disparity map for each frame to describe the scene's depth distribution. As numerous studies (Sindagi et al. 2020; Yang et al. 2022; Li et al. 2023; Yan et al. 2021; Zhou et al. 2023) have suggested, depth distribution is crucial for vision tasks under adverse weather conditions. For example, raindrops on the windshield have a small depth, and the rain veiling effect is correlated with the depth distribution.

We collected over 30 videos from July to December 2023. By comparing these recordings with data from the weather stations, we identified four videos captured under violent rainfall conditions, with hourly rainfall intensities of 63.0 mm, 76.0 mm, 64.8 mm, and 56.8 mm, respectively. For each video, we manually checked and removed redundant clips caused by waiting at traffic lights or traffic jams.

**Privacy Protection** Our project is supervised by the Institutional Review Board (IRB). To ensure privacy protection, we reported all our plans and activities to the IRB before and during the data collection and annotation stages. In compliance with their requirements, we removed all video sequences containing private properties and masked all human faces and car plates. This careful adherence to privacy guidelines underscores our commitment to safeguarding personal information throughout the project. After these processes, we selected 14,757 unlabeled frames and 100 labeled frames to construct our dataset.

**Annotation Procedure** We strictly adhere to the well-recognized Cityscapes annotation guidelines and the 19 classes protocol. However, since trains are not encountered in driving in Singapore, our videos do not include any trains. As a result, the 'train' class from the 19 classes protocol is excluded, leaving us with a total of 18 classes. When selecting frames to label, we annotate every 21st frame. This approach allows video semantic segmentation experiments to gather temporal information from either the 20 frames preceding the current frame or the 20 frames following it.

Accurately annotating a single image under violent rainfall conditions is challenging, even for human annotators. However, the videos in our dataset provide the advantage of utilizing temporal information from adjacent frame sequences, aiding in more accurate annotations compared to relying on a single image. To further minimize potential manual labeling errors, we require at least two members of the annotation team to peer-review each annotation.

## Comparison to Related Datasets

We compare our dataset against existing rain datasets from two perspectives: degradation levels and parameters.

**Degradation Levels** Although all rainy images are affected by degradations, the types and severity of these degradations can vary with different rainfall intensities. Under light rainy conditions, the degradations are marginal. As shown in Fig. 2 (a), we can observe only wet floors and sparse raindrops on the windshield.

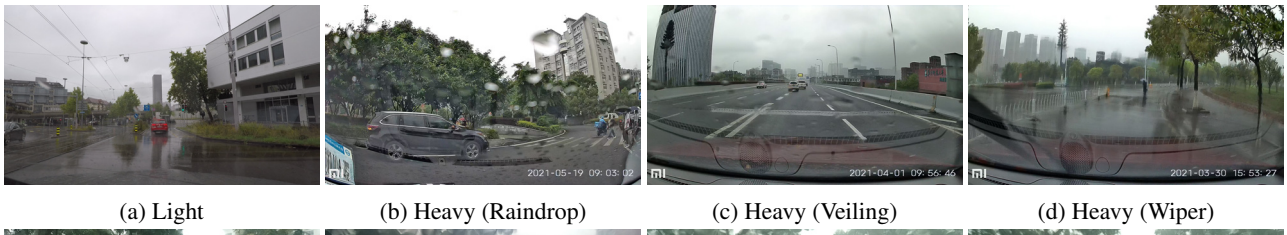
As rainfall intensity escalates to heavy rain, both raindrop and rain veiling effect become markedly evident. Firstly, raindrops increase in size and density. However, as demonstrated by Fig. 2 (b) in the heavy rain dataset (Zhong et al. 2022), these raindrops retain clear boundaries. These clear boundaries aid existing models in detecting raindrops, thus simplifying the challenge, as illustrated in Figs. 3 (a-b). Secondly, a fog-like rain veiling effect emerges due to the accumulation of rain streaks. This veiling effect is more pronounced in areas farther from the camera, impairing visual clarity, as evidenced by the distant 'building' in Figs. 2 (c-d). Additionally, under heavy rain conditions, the use of wipers introduces further visual obstructions. The wipers cannot completely eliminate the raindrops, leaving a thin layer of water residue that blurs the wiped areas, as depicted in Fig.2 (d) and discussed in (Zhong et al. 2022).

Under violent rainfall conditions, excessive large raindrops easily form water flows, as illustrated in Fig. 2 (e). In contrast to the large raindrops in heavy rain, the water flows in our dataset lack clear boundaries and cannot be identified by existing models, as shown in Figs. 3 (c-d). The uneven distribution of water flow thickness leads to varying refraction factors across the affected area, causing extensive refraction and distortion. Secondly, the increased intensity of rain streaks intensifies the rain veiling effect, as observed on 'vegetation' and 'building' in Fig. 2 (f). Additionally, the increased rainfall intensity necessitates more frequent use of wipers, exacerbating the visual degradation caused by their motion as explained and presented in Figs. 2 (g-h).

**Parameters** We compare different parameters of our dataset with other rainy datasets, as shown in Tab. 1. Our dataset has a relatively high resolution, enabling models to learn more fine-grained semantic segmentation predictions. With a total of 14,857 frames, it ranks as the third-largest rainfall dataset available. Annotating under severe weather degradation is both challenging and costly, even for human annotators, but we have successfully provided a comparable number of labeled frames. Our dataset features real-world urban scenes and is one of only three datasets that support video-based methods. Additionally, it is the only real-world rainy dataset that includes disparity information. Most notably, our dataset uniquely captures violent rainfall conditions, distinguishing it from existing datasets.

# Experiments

In this section, we benchmark four common models or methods to evaluate their robustness under violent rainfall conditions. For all experiments, we use the mean Intersection over Union (mIoU) percentage as our evaluation metric, where a



(a) Light (b) Heavy (Raindrop) (c) Heavy (Veiling) (d) Heavy (Wiper)

(e) **Violent (Water flow)** (f) **Violent (Veiling)** (g) **Violent (Wiper before)** (h) **Violent (Wiper after)**

Figure 2: Degradations under different rainfall intensities. **Light:** In (a), we can observe that the visual is not obviously impaired. **Heavy:** In (b), large and dense raindrops refract the surrounding environment. In (c) and (d), distant buildings are veiled by a fog-like accumulation of rain. In (d), after the wiper has passed, raindrops are wiped into a thin layer of water, blurring the scene. **Violent:** In (e), numerous raindrops have merged to form water flows that obscure the entire windshield, significantly impairing visibility. In (f), the 'vegetation' and 'building' are significantly affected by the rain veiling effect, despite being closer than the 'building' in (c) and (d). (g-h) show frames before and after the wiper is applied. Before the wiper, the detrimental water flow corrupts most visual clues, making it impossible to distinguish the 'road', 'sidewalk', 'fence', and 'building'. After the wiper, the scene remains blurred, and due to the high rainfall intensity, new raindrops quickly land in the blurred area, exacerbating the degradation. Zoom in for a better visibility.

(a) Heavy (Raindrop) (b) Segmentation Results (c) **Violent (Water flow)** (d) **Segmentation Results**

Figure 3: Segmentation results from SAM on heavy rainfall intensity (Zhong et al. 2022) and violent rainfall intensity. In (b), we observe that SAM can identify large and dense raindrops with clear boundaries. However, in (d), SAM fails to identify the water flow. Additionally, in certain areas, such as the 'terrain' and parts of the 'building', SAM struggles to interpret the degraded scene and consequently does not generate any masks. Zoom in for a better visibility.

| Dataset | Resolution | Label/Total | Real | Video | Disparity | Intensity |
|---|---|---|---|---|---|---|
| Raincouver | 1,280×720 | 285/326 | ✓ | × | × | Light |
| nuScenes | 1,600×900 | 58/1,300 | ✓ | × | × | Light |
| BDD100K (Rain) | 1,280×720 | 253/5,808 | ✓ | × | × | Light |
| ACDC (Rain) (Sakaridis, Dai, and Van Gool 2021) | 1,920×1,080 | 627/1,000 | ✓ | × | × | Heavy |
| RainCityscapes | 2,048×1,024 | 1,760/1,620 | × | × | ✓ | Heavy |
| RaidaR (Jin et al. 2021) | 1,920×1,280 | 5,000/58,452 | ✓ | × | × | Light |
| Rainy WCity | 1,920×1,080 | 500/24,335 | ✓ | × | × | Heavy |
| MVSS (Rain) (Ji et al. 2023) | 630×460 | 6/732 | ✓ | × | × | Light |
| ERF (Ours) | 1,920×1,080 | 100/14,857 | ✓ | ✓ | ✓ | Violent |

Table 1: Comparison of rainy datasets used for semantic segmentation task.

higher value indicates more accurate semantic segmentation map predictions.

## Image-Based Foundation Models

Image-based foundation models are large-scale neural networks trained on vast datasets of images to understand and generate visual content. These models offer significant ad-

vantages over traditional methods due to their extensive pre-training on diverse datasets, which enables them to capture complex visual patterns and generalize well across various tasks and conditions. This generalization ability is particularly crucial for adverse conditions, where they maintain high performance due to their exposure to diverse and challenging scenarios during pre-training, allowing them to adapt to new, unseen situations.

We evaluate four foundation models: CLIP'21 (Radford et al. 2021), SAM'23 (Kirillov et al. 2023), EVA02'23 (Fang et al. 2023) and DINOv2'23 (Oquab et al. 2023). In our experiments, these models are kept frozen as encoders, and we use the Mask2Former decoder (Cheng et al. 2022) for all the models, as suggested in (Wei et al. 2024). We fine-tune the decoders for semantic segmentation tasks on the Cityscapes dataset. We follow the exact training configurations suggested in these papers to ensure a fair comparison. The performance are presented in Tab. 2 under foundation models. We observe that DINOv2 achieves the best performance, possibly due to its larger pretraining dataset and model size compared to the other models. Additionally, their augmented and curated data preprocessing stage likely enhances the model's robustness to unseen domains.

## Image-Based Domain Generalization Methods

Domain generalization methods enhance model robustness by learning domain-invariant features, enabling models to perform well on unseen domains without needing target domain data for fine-tuning. For instance, a model trained to perform semantic segmentation in rainy conditions learns to focus on essential features like the shapes and contours of objects rather than rain degradations, making it effective in new, unseen scenarios. An advantage of this method is that it can integrate with various types of models, including foundation models.

We evaluate four domain generalization methods: HRDA'22 (Hoyer, Dai, and Van Gool 2022b), HGFormer'23 (Ding et al. 2023), PASTA'23 (Chattopadhyay et al. 2023) and Rein'24 (Wei et al. 2024). For these methods, we use their recommended encoders and decoders. HRDA employs a transformer-based encoder, MiT-B5 (Xie et al. 2021), and a DAFormer (Hoyer, Dai, and Van Gool 2022a) decoder. HGFormer uses a transformer-based encoder, Swin-L (Liu et al. 2021), and their own proposed decoder. The latest methods, PASTA and Rein, can already integrate with the foundation model DINOv2, with a Mask2Former decoder. Following their recommendations, we train these models on the Cityscapes dataset and adhere to the suggested training configurations for all models. The performance are presented in Tab. 2 under domain generalization methods. We observe that the foundation model-based methods, PASTA and Rein, significantly outperformed the other methods, highlighting the importance of incorporating foundation models for handling images under adverse weather conditions. Additionally, Rein achieved a 3.6 mIoU (%) performance gain compared to the second-best method, PASTA. A possible reason for this difference is that PASTA is an augmentation-based, model-agnostic method designed for generalizing across different models, integrating with a frozen DINOv2 without any refinement. In contrast, Rein is specifically designed for foundation models and refines DINOv2 during the training process, leading to better integrated performance.

## Image-Based Domain Adaptation Methods

When there is a discrepancy between the training source data and the testing target data, domain adaptation methods can be applied to learn to minimize the discrepancies and enhance the performance of machine learning models. For example, a model trained on ideal weather images (source) may struggle with rainy scenes (target); domain adaptation learns and bridges this gap. Unlike foundation models, which rely on extensive pre-training on diverse datasets, and domain generalization methods, which aim to perform well across various domains without specific adaptation, domain adaptation specifically fine-tunes models for rainy scenarios. This fine-tuning is achieved through techniques such as adversarial training, where the model learns to minimize the differences between the source and target domains, feature alignment, where features from both domains are mapped to a common space, or self-learning, where models are updated based on pseudo-labels generated from teacher models. This makes domain adaptation especially advantageous for our task, ensuring that models remain robust and accurate in the target rainy environment.

We evaluate four domain adaptation methods: DAFormer'22 (Hoyer, Dai, and Van Gool 2022a), HRDA'22 (Hoyer, Dai, and Van Gool 2022b), MIC'23 (Hoyer et al. 2023) and CDAC'23 (Wang et al. 2023). For these methods, we use their recommended encoder, MiT-B5, and the DAFormer decoder. Based on their recommendations, we use the Cityscapes as the source, and our dataset as the target. In the training, we strictly follow their suggested training configurations.

The performance results are presented in Tab. 2 under domain adaptation methods, where we observe that HRDA demonstrates the best performance. However, compared to foundation models and domain generalization methods, domain adaptation methods are less effective under violent rainfall conditions. There might be two reasons: (1) Domain adaptation requires to learn the discrepancies between the source and target domains. When these discrepancies are too large and complex, existing methods struggle to effectively learn and reduce them, resulting in less robust models for the target condition. (2) As discussed in the domain generalization section, the involvement of foundation models is crucial and can provide significant improvement. However, the existing domain adaptation methods have not yet integrated with such models to further enhance their performance.

Compared to HRDA in the domain generalization setting, we observe that HRDA in the domain adaptation setting shows an improvement of 2.4 mIoU (%). The only difference between these two models is that, in the domain adaptation setting, the model additionally receives the unlabeled target violent rainfall images as additional inputs. This improvement indicates that incorporating unlabeled images for fine-tuning can enhance the model's robustness against violent rainfall conditions. Therefore, when combined properly with foundation models, domain adaptation methods have

| Method | Road | Side W. | Build. | Wall | Fence | Pole | Traf.L. | Sign | Vege. | Terrain | Sky | Person | Rider | Car | Truck | Bus | Motor. | Bike | mIoU↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Foundation Models*** | | | | | | | | | | | | | | | | | | | |
| CLIP | 91.0 | 38.2 | 70.6 | 36.0 | 47.6 | 44.6 | 67.2 | 55.5 | 87.3 | 71.7 | 91.7 | 38.2 | 13.9 | 83.2 | 9.1 | 58.6 | 38.1 | 9.3 | 52.5 |
| SAM | 82.6 | 25.7 | 68.5 | 47.0 | 63.8 | 48.4 | 70.8 | 57.2 | 87.9 | 71.8 | 89.4 | 37.9 | 15.8 | 83.9 | 12.0 | 57.2 | 20.7 | 1.4 | 52.0 |
| EVA02 | 91.0 | 40.6 | 70.9 | 35.0 | 55.9 | 44.2 | **71.1** | 56.1 | 87.0 | 72.0 | 90.2 | 41.0 | **21.8** | 85.6 | 9.8 | 64.1 | 41.3 | 9.1 | 54.4 |
| DINOv2 | **93.4** | **51.4** | **73.0** | **50.6** | **69.8** | **54.0** | 56.3 | **60.3** | **88.9** | **77.9** | **92.0** | **44.9** | 17.4 | **88.8** | **18.3** | **79.8** | **48.1** | 15.4 | **59.8** |
| ***Domain Generalization Methods*** | | | | | | | | | | | | | | | | | | | |
| HRDA | 88.7 | 31.8 | 65.9 | 24.5 | 57.3 | 34.0 | 57.7 | 45.3 | 87.7 | 70.8 | 89.4 | 34.6 | 13.2 | 72.4 | 2.4 | 46.2 | 38.7 | 3.2 | 47.6 |
| HGFormer | 90.8 | 34.1 | 68.1 | 29.3 | 49.7 | 45.9 | **77.4** | 49.7 | 86.9 | 69.7 | 91.1 | 47.1 | 13.1 | 77.4 | 7.6 | 57.5 | 45.2 | 22.2 | 53.0 |
| PASTA | 93.0 | **53.1** | 73.8 | 46.9 | 65.8 | 54.0 | 57.0 | 60.7 | **88.7** | **77.5** | **91.3** | 38.9 | **19.6** | 88.7 | 13.7 | 80.3 | 48.1 | 11.7 | 58.8 |
| Rein | **93.3** | 49.7 | **73.9** | **53.9** | **70.3** | **57.2** | 60.1 | **61.1** | 88.3 | 76.6 | 91.0 | **48.1** | 18.04 | **89.9** | **34.5** | **87.2** | **48.4** | **25.3** | **62.4** |
| ***Domain Adaptation Methods*** | | | | | | | | | | | | | | | | | | | |
| DAFormer | 57.6 | 21.4 | 65.2 | 39.9 | 56.2 | 42.1 | 44.6 | 51.1 | 85.6 | 44.0 | 53.0 | 20.0 | 11.1 | 67.4 | 0.1 | 29.4 | **47.6** | 10.0 | 41.2 |
| HRDA | 70.8 | 33.5 | 71.7 | 46.3 | **68.5** | 49.7 | 54.0 | 60.5 | 88.9 | **74.9** | 70.4 | 45.6 | **13.3** | 70.4 | 0 | 53.6 | 19.3 | **15.9** | **50.0** |
| MIC | 81.1 | 17.2 | **75.9** | **49.3** | 68.2 | **50.2** | **55.5** | **62.6** | **89.9** | 17.3 | 80.1 | **51.9** | 12.0 | **79.9** | **7.1** | **70.6** | 28.8 | 8.0 | 49.9 |
| CADC | **84.6** | **39.5** | 63.0 | 42.2 | 61.6 | 37.7 | 52.1 | 57.5 | 86.3 | 71.0 | **86.3** | 31.1 | 9.9 | 78.3 | 0 | 47.3 | 33.2 | 2.2 | 48.8 |

Table 2: Semantic segmentation performance (IoU in %) on four foundation models, four domain generalization methods, and four domain adaptation methods. **Bold** numbers indicating the best scores in each category.

the potential to achieve superior performance compared to foundation model-based domain generalization methods.

## Video-Based Methods

Video-based methods extend the principles of image-based methods to the temporal domain, making predictions based on video sequences. Unlike image-based methods, which focus on single frames, video-based adaptation leverages temporal information and motion cues across consecutive frames. This approach captures the dynamics of adverse weather conditions more effectively, such as changes in raindrops or motion of water flow over time. This approach more effectively captures the dynamics of adverse weather conditions, such as changes in raindrops or the motion of water flow over time. By considering the changing degradations in the temporal domain, video-based methods can reduce these degradations and more accurately identify the true scene distorted by them.

We evaluate three video-based methods: DA-VSN'21 (Guan et al. 2021), TPS'22 (Xing et al. 2022), and SFC'23 (Gao et al. 2023). Unlike image-based methods and models, video-based methods have specific requirements for inference speed. To achieve an optimal accuracy-speed tradeoff, these video-based methods use a lightweight backbone, Accel (Jain, Wang, and Gonzalez 2019), which includes an optical flow estimation module, a fusion layer, and a semantic segmentation framework, DeeplabV2 (Chen et al. 2017). In our experiments, we strictly adhere to the configurations suggested in these papers, using the same backbone and training setups. Since these methods are all domain adaptation-based techniques, they require a source dataset. As recommended, we use the urban scene, ideal weather video dataset, Viper (Richter, Hayder, and Koltun 2017), as the source, and our dataset as the target. Since Viper does not follow the Cityscapes 19 class protocols, we evaluate on the 15 common classes, following the same settings as in (Guan et al. 2021; Xing et al. 2022; Gao et al. 2023).

The performance results are presented in Tab. 3, where we observe a significant drop in performance for video-based methods compared to image-based models and methods, despite the incorporation of additional temporal information. This suggests there is potential for improving the current backbone to achieve more resilient and accurate performance in the challenging environment of violent rainfall.

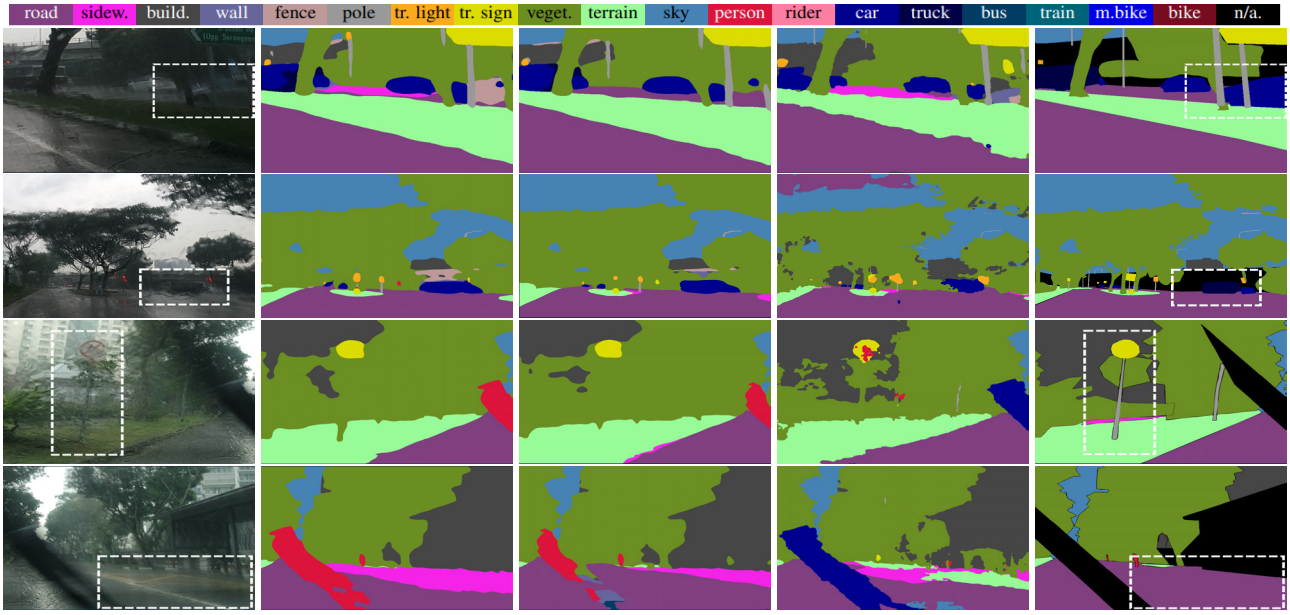
## Qualitative Result

We present the qualitative results of DINOv2, Rein, and HRDA (domain adaptation) in Fig. 4. These models were chosen for their top performance within their respective categories. Video-based methods are not included in this comparison due to their comparatively lower performance.

In the first row, we demonstrate how water flow degradation can affect these models. In this row, the car in the white box is blocked by water flow, completely blurred, and corrupted by the refracted information from the surrounding environment. This misleads DINOv2 to classify it as 'fence', Rein to classify it as 'road', and HRDA to classify it as 'wall'. In the second row, the distant 'bus' and 'car' in the white box are affected by both rain veiling and water flow. As a consequence, all the models failed to accurately segment them and misclassified the two different vehicles as 'car'. In the last two rows, we investigate how the wipers

| Method | Road | S.walk | Build. | Fence | Tr.Light | Sign | Veget. | Terrain | Sky | Person | Car | Truck | Bus | M.bike | Bike | mIoU↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DA-VSN | 44.7 | 8.3 | 44.7 | 23.8 | 4.0 | 24.1 | 73.3 | 21.0 | 72.1 | 5.4 | **68.0** | 0.6 | 11.6 | 0.2 | 0 | 26.8 |
| TPS | 34.7 | 7.7 | 49.7 | **34.4** | 2.7 | 25.0 | 75.5 | 20.5 | 80.3 | 8.2 | 62.6 | 1.8 | **25.5** | 0 | **0.8** | 28.6 |
| SFC | **82.5** | **11.5** | **60.0** | 14.6 | **38.3** | **30.7** | **81.0** | **46.3** | **88.9** | **9.7** | 66.9 | **2.2** | 0 | **14.2** | 0 | **36.5** |

Table 3: Semantic segmentation performance (IoU in %) on three different video-based methods.




(a) Image (b) DINOv2 (c) Rein (d) HRDA (e) Ground Truth

Figure 4: Comparisons of semantic segmentation performance using the foundation model DINOv2, the domain generalization method Rein, the domain adaptation method HRDA, and ground truths on the ERF dataset. Different colors represent different classes. The 'n/a' areas, such as the wiper and the viaduct, are ignored when computing mIoU.

can degrade the scene and mislead the models. These two rows are from the same image, where the fourth row shows the part before the wiper, and the fifth row shows the part after the wiper. Before the wiper passes, we see severe water flow degradation, causing all the models to fail in identifying the 'pole'. After the wiper passes, although the water flow degradation is significantly reduced, there is still blurriness from a thin layer of water residue and small raindrops in the white box area. As a result, all the models misclassified the 'road' as either 'sidewalk' or 'terrain'.

Qualitatively, we observe that foundation model-based methods are more robust to violent rainfall conditions. In contrast, HRDA (domain adaptation) exhibits some obvious hallucinations (such as in the sky area in the third row) and produces jagged, rough segmentation.

## Discussion and Suggestions

In this study, we propose the Extreme RainFall (ERF) dataset, the first dataset under violent rainfall conditions, specifically designed for both image and video semantic segmentation tasks. We benchmarked the robustness of various models and methods using our dataset. Our evaluations focused on four categories: image-based foundation models, image-based domain generalization methods, domain adaptation methods, and video-based methods. The models integrated with foundation models demonstrated a clear advantage over other models under violent rainfall conditions. Additionally, the domain generalization method, Rein, further improved the performance of DINOv2, indicating that domain generalization methods remain effective under these conditions. By comparing HRDA in both domain generalization and domain adaptation settings, we observed that, despite the significant domain gap between the source (ideal weather conditions) and the target (violent rainfall conditions), involving the unlabeled target dataset can enhance the model's robustness. Thus, developing a foundation model-based domain adaptation method, similar to Rein, could be a promising solution for extreme rainfall events. For video-based methods, their current performance does not match that of image-based methods. Given the dynamic nature of rainfall degradation, incorporating temporal information is a potential solution. However, to achieve this, further research is needed to improve the existing video-based methods' backbones, effectively incorporate temporal information, and maintain comparable inference speeds.

## Acknowledgements

We would like to express our gratitude to Yahong Jia and to Bindong Shi for their significant contributions to data collection. We are also deeply thankful to Prof. Robby T. Tan for his invaluable guidance and support throughout this work. This project is supported by the National Research Foundation Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation.

## References

Bank, W. 2024. World Development Indicators. Accessed: 2024-06-02.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Chattopadhyay, P.; Sarangmath, K.; Vijaykumar, V.; and Hoffman, J. 2023. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19288–19300.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Chen, S.; Ye, T.; Zhang, K.; Xing, Z.; Lin, Y.; and Zhu, L. 2025. Teaching Tailored to Talent: Adverse Weather Restoration via Prompt Pool and Depth-Anything Constraint. In *European Conference on Computer Vision*, 95–115. Springer.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

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.

Ding, J.; Xue, N.; Xia, G.-S.; Schiele, B.; and Dai, D. 2023. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15413–15423.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Gao, Y.; Wang, Z.; Zhuang, J.; Zhang, Y.; and Li, J. 2023. Exploit domain-robust optical flow in domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 641–649.

Guan, D.; Huang, J.; Xiao, A.; and Lu, S. 2021. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8053–8064.

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 *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 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.

Hu, X.; Fu, C.-W.; Zhu, L.; and Heng, P.-A. 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8022–8031.

Jain, S.; Wang, X.; and Gonzalez, J. E. 2019. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8866–8875.

Ji, W.; Li, J.; Bian, C.; Zhou, Z.; Zhao, J.; Yuille, A. L.; and Cheng, L. 2023. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1094–1104.

Jin, J.; Fatemi, A.; Lira, W. M. P.; Yu, F.; Leng, B.; Ma, R.; Mahdavi-Amiri, A.; and Zhang, H. 2021. Raidar: A rich annotated image dataset of rainy street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2951–2961.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, M.; Xie, B.; Li, S.; Liu, C. H.; and Cheng, X. 2023. VBLC: visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8605–8613.

Li, X.; Kou, K.; and Zhao, B. 2021. Weather GAN: Multi-domain weather translation using generative adversarial networks. *arXiv preprint arXiv:2103.05422*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Özdenizci, O.; and Legenstein, R. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10346–10357.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Richter, S. R.; Hayder, Z.; and Koltun, V. 2017. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2213–2222.

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.

Sindagi, V. A.; Oza, P.; Yasarla, R.; and Patel, V. M. 2020. Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 763–780. Springer.

Singapore, M. S. 2024. Climate Historical Daily. Accessed: 2024-06-02.

Tung, F.; Chen, J.; Meng, L.; and Little, J. J. 2017. The raincouver scene parsing benchmark for self-driving in adverse weather and at night. *IEEE Robotics and Automation Letters*, 2(4): 2188–2193.

Wang, K.; Kim, D.; Feris, R.; and Betke, M. 2023. CDAC: Cross-domain Attention Consistency in Transformer for Domain Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11519–11529.

Wei, Z.; Chen, L.; Jin, Y.; Ma, X.; Liu, T.; Lin, P.; Wang, B.; Chen, H.; and Zheng, J. 2024. Stronger, fewer, & superior: Harnessing vision foundation models for domain generalized semantic segmentation.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xing, Y.; Guan, D.; Huang, J.; and Lu, S. 2022. Domain adaptive video segmentation via temporal pseudo supervision. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 621–639. Springer.

Yan, W.; Tan, R. T.; Yang, W.; and Dai, D. 2021. Self-aligned video deraining with transmission-depth consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11966–11976.

Yang, X.; Mi, M. B.; Yuan, Y.; Wang, X.; and Tan, R. T. 2022. Object detection in foggy scenes by embedding depth and reconstruction into domain adaptation. In *Proceedings of the Asian Conference on Computer Vision*, 1093–1108.

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.

Zhong, X.; Tu, S.; Ma, X.; Jiang, K.; Huang, W.; and Wang, Z. 2022. Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding. In *IJCAI*, 1743–1749.

Zhou, H.; Chang, Y.; Yan, W.; and Yan, L. 2023. Unsupervised Cumulative Domain Adaptation for Foggy Scene Optical Flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9569–9578.