Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding

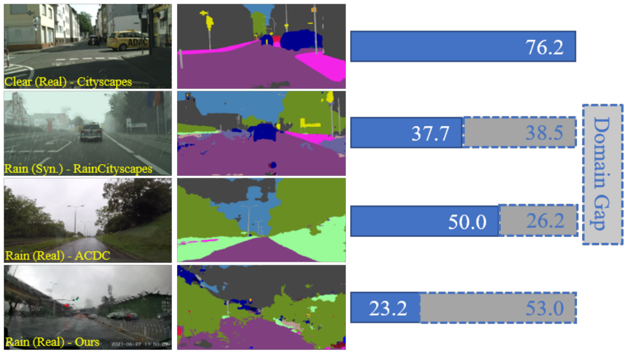
Xian Zhong[1,2,*] , **Shidong Tu**[1] , **Xianzheng Ma**[3,*] , **Kui Jiang**[3,†] , **Wenxin Huang**[4] and **Zheng Wang**[3,†]
[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology
[2] School of Electronic Engineering and Computer Science, Peking University
[3] School of Computer Science, Wuhan University
[4] School of Computer Science and Information Engineering, Hubei University
zhongx@whut.edu.cn, neowell@126.com, {maxianzheng, kuijiang}@whu.edu.cn,
wenxinhuang_wh@163.com, wangzwhu@whu.edu.cn

Abstract

Scene understanding in adverse weather conditions (*e.g.*, rainy and foggy days) has drawn increasing attention, arising some specific benchmarks and algorithms. However, scene segmentation under rainy weather is still challenging and under-explored due to the following limitations on the datasets and methods: 1) Manually synthetic rainy samples with empirically settings and human subjective assumptions; 2) Limited rainy conditions, including the rain patterns, intensity, and degradation factors; 3) Separated training manners for image deraining and semantic segmentation. To break these limitations, we pioneer a real, comprehensive, and well-annotated scene understanding dataset under rainy weather, named **Rainy WCity**. It covers various rain patterns and their bring-in negative visual effects, covering wiper, droplet, reflection, refraction, shadow, windshield-blurring, *etc*. In addition, to alleviate dependence on paired training samples, we design an unsupervised contrastive learning network for real image deraining and the final rainy scene semantic segmentation via multi-task joint optimization. A comprehensive comparison analysis is also provided, which shows that scene understanding in rainy weather is a largely open problem. Finally, we summarize our general observations, identify open research challenges, and point out future directions.

1 Introduction

Out-door application, *e.g.*, semantic segmentation under the driving situation, has achieved impressive progress with the existing segmentation models under clear weather conditions. However, the performance of these methods drops sharply when facing inclement weather, since adverse weather can significantly degrade the image quality and readability. One typical scenario is the rain condition, as shown in Fig. 1, with multiple degradation scenes and rain patterns, making the segmentation task more challenging.



(a) Input (b) Prediction (c) mIoU

Figure 1: Column (a) from top to bottom: Clear image from Cityscapes, synthetic rainy images from RainCityscapes, real rainy images respectively from ACDC and our Rainy WCity. Column (b) presents their segmentation results by DeepLabV3+. Due to the limited rain conditions and degradation factors, their rainy day scenarios and predictions are closer to clear days. However, when all these factors are included, it is almost impossible for the model to generate discernible results. Column (c) shows the mean Intersection over Union (mIoU) on the validation part of these datasets by DeepLabV3+. The gray bar shows the largest domain gap between real rainy images from our Rainy WCity and clear images from Cityscapes, compared to the gap between other real rainy dataset (ACDC)/synthetic dataset (RainCityscapes) and Cityscapes.

To promote this task, there are at least two urgent questions required to be addressed: how to construct a benchmark for comprehensively evaluating and training the segmentation model under rainy scenes? And how to make the segmentation model robust to diverse rainy scenes?

For the first question, some researchers recently put their efforts towards dataset generation and collection, including synthetic and real rainy datasets. Some studies aggregate the depth information [Hu *et al.*, 2021] to generate the rainy samples or spray water on the glass to simulate driving on rainy days [Porav *et al.*, 2020]. Although these schemes have boosted the development of segmentation tasks under driving scenes, there still exists significant distribution discrepancy in the real rainy scenes, and the synthetic data lose the randomness that may exist. In particular, the data

[*] These authors contributed equally to this work.
[†] Corresponding authors.

Dataset	Resolution	Label/Total	Real	Scenario					Intensity		
				Occlusion	Blur	Droplet	Reflection	Wiper	Light	Moderate	Heavy
Cityscapes [Cordts *et al.*, 2016]	2,048×1,024	0/0	×	×	×	×	×	×	×	×	×
Raincouver [Tung *et al.*, 2017]	1,280×720	285/326	✓	×	✓	✓	✓	×	✓	×	×
KITTI [Alhaija *et al.*, 2018]	1,382×512	0/0	×	×	×	×	×	×	×	×	×
RID [Li *et al.*, 2019]	Variable	0/2,495	✓	✓	✓	✓	✓	×	✓	×	×
Apolloscape [Huang *et al.*, 2020]	3,384×2,710	0/0	×	×	×	×	×	×	×	×	×
nuImages [Caesar *et al.*, 2020]	1,600×900	58/1,300	✓	×	✓	×	✓	×	✓	×	×
BDD [Yu *et al.*, 2020]	1,280×720	253/5,808	✓	✓	✓	✓	✓	×	✓	×	×
ACDC [Sakaridis *et al.*, 2021]	1,920×1,080	1,000/1,000	✓	✓	✓	×	✓	✓	✓	×	×
RainCityscapes [Hu *et al.*, 2021]	2,048×1,024	1,760/10,620	×	×	✓	✓	×	×	✓	✓	✓
RaidaR [Jin *et al.*, 2021]	1,920×1,080	5,000/58,542	✓	✓	✓	×	✓	×	✓	×	×
Rainy Wcity (Ours)	1,920×1,080	500/24,335	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of driving datasets in terms of the sample resolution, annotated rainy images, total rainy images, authenticity of rain, scene diversity, and rain patterns, respectively. "Reflection" means road surface reflection due to rain droplet accumulation. "Wiper" means the effects on the segmentation model due to the movements of wipers. While researchers have previously focused on introducing the essential characteristics of rainy days into their experiments, it is time to take it a step further to better cope with complex scenarios in practice.

discrepancy arises a problem that the model trained by the synthetic data losses the ability sharply in actual scenarios. For the existing real rainy datasets [Jin *et al.*, 2021; Sakaridis *et al.*, 2021], samples are closer to post-rain scenes since the characteristics of rainy days have not been included, *e.g.*, raindrops, road reflection, and droplet occlusion.

As shown in Fig. 1, the representative segmentation algorithm DeepLabV3+ [Chen *et al.*, 2018] (trained on the Cityscapes dataset) shows better performance on the clear Cityscapes or synthetic rainy Cityscapes scenarios. In contrast, the segmentation performance declines precipitously in the real-world rainy scenes, particularly in our dataset (from 76.2% to 23.2% of the mIoU), since there are more complex degradation factors.

To break these limitations aforementioned, we construct a novel rainy dataset with accurate semantic annotations, named Rainy WCity, to investigate the segmentation task for driving scenes comprehensively. Rainy WCity covers 5 common driving scenes on rainy days with a total of 24,335 real rainy samples, where each scene has 100 images with the corresponding well-annotation segmentation label.

Our dataset distinguishes others by featuring the following elements: **1) Diverse Rain Patterns:** including light, moderate, and heavy rain scenarios; **2) Diverse degradation factors:** besides the typical rain occlusion, our dataset considers numerous adverse effects of degradation, commonly occurred on rainy days but overlooked in other real rainy datasets, *e.g.*, the droplet interference, road reflection, and windshield-blurring. In particular, we provide the elaborated classification of these scenes with the corresponding well-annotated segmentation labels.

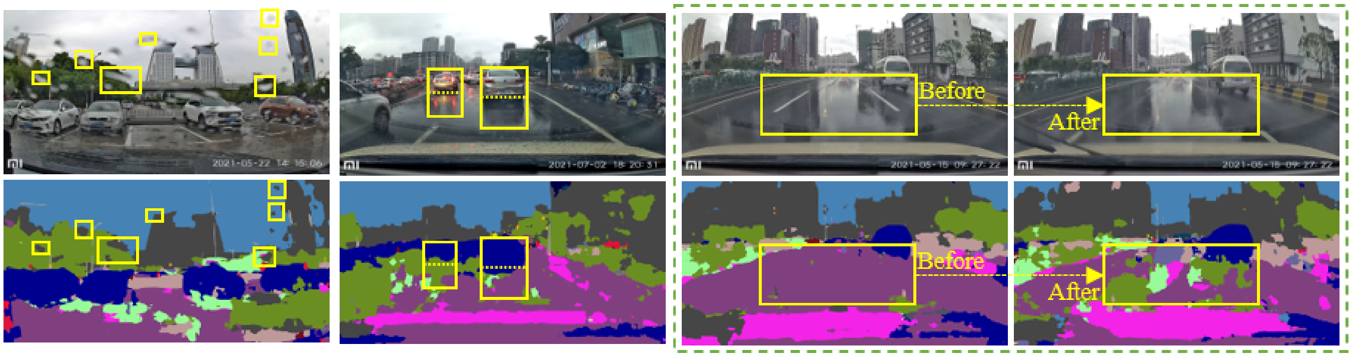
For the second question, researchers have developed efforts to design the two-stage optimization model by simply cascading the image deraining and semantic segmentation tasks for the rainy segmentation [Jiang *et al.*, 2020]. As a result, the segmentation results are barely satisfactory in real challenging scenarios, while the marginal improvement comes at the cost of extra computation and memory usage. The reason lies in that these two tasks have a significant discrepancy with regard to the optimization objective. The former tends to learn the pixel-wise fidelity, however, the latter focuses on the semantic-wise fidelity. Unlike these methods, we propose to achieve the joint training of the image deraining and semantic segmentation and construct an unsupervised rainy scene segmentation method. In particular, we introduce scene segmentation that meets real rain (S2R2). In addition, to eliminate the dependence on the real paired training samples, we train the deraining network via unsupervised contrastive learning. Meanwhile, the segmentation counterpart is optimized via cross-entropy loss. Experimental results on Rainy WCity demonstrate that our proposed S2R2 method achieves appealing improvements over the state-of-the-art methods in terms of deraining performance (image clarity) and segmentation accuracy.

Overall, contributions can be summarized threefold:

- We re-investigate the segmentation task under real rainy scenes and construct a real and comprehensive dataset Rainy WCity with well annotations. It provides a fair and unified benchmark for further research on this challenging task.
- We design an unsupervised joint optimization framework (deraining and semantic segmentation) to significantly promote segmentation performance in rainy scenarios, serving as a baseline method for further research.
- We conduct comprehensive experimental comparison under diverse rainy scenarios and provide thoughtful insights and analysis on this task, which is momentous for opening research and future direction.

2 Rainy WCity Dataset

This section details the difference between our proposed dataset and the existing driving datasets, including Cityscapes [Cordts *et al.*, 2016], Raincouver [Tung *et al.*, 2017], KITTI [Alhaija *et al.*, 2018], RID [Li *et al.*, 2019], Apolloscape [Huang *et al.*, 2020], nuImages [Caesar *et al.*, 2020], BDD [Yu *et al.*, 2020], ACDC [Sakaridis *et al.*, 2021], RainCityscapes [Hu *et al.*, 2021], and RaidaR [Jin *et al.*, 2021]. Table 1 provides the comparison results in terms of the image resolution, sample number, scenarios, and rain patterns. Our dataset provides a further differentiation of some degradation factors under real rainy scenes.



(a) Droplet (b) Reflection (c) Before wiper (c) After wiper

Figure 2: Segmentation results by DeepLabV3+ under different rainy driving situations. (a) shows the effects of the droplet. Multiple locations covered by droplets are misidentified. (b) shows the effect of road reflection. Here, the shadow of cars is incorrectly identified. (c) shows the blur effect of the wiper. When the wiper scrapes across the glass, although there seems to be no problem in the human eye, the actual segmentation effect significantly deviates.

2.1 Diverse Degradation Scenarios

Shading Effect of Droplet Droplet, the most common phenomenon on rainy days, will prevent light from passing straight through the object, and when it falls on the windshield, the content distortion caused by the reflectance and refraction becomes more pronounced. In Fig. 2(a), areas covered with rain droplet has significant misidentification, *e.g.*, the green "vegetation" is identified as the grey "building", while the blue "sky" is misidentified as the grey "building".

Road Reflection The rain droplets accumulated on the road commonly form a mirror reflection effect, confusing the segmentation model with the fake and indistinguishable objective boundary. As shown in Fig. 2(b), the shadow of cars is incorrectly identified.

Blurring by Windshield Wiper The wiper eliminates the visual occlusion caused by the rain accumulation. However, for the real rainy days, the wiper operation could evidently promote the visual clarity of human observed views, but the unexpected change of scene contents is unfriendly to the segmentation model for the real-time application. As is shown in the yellow box in Fig. 2(c), the result is not consistent with the human's view.

2.2 Collection

Data collection starts at the end of March 2021 and has lasted over five months in total. We mount four cameras on four cars respectively on their roof behind the windshield. All four employed cameras are the same model: Xiaomi Recorder 2, which enables recording 30 frames per *second* with the resolution of 1,920×1,080. We have collected 5,619 videos in total and coarsely removed the rain-less and night parts, resulting in 1,685 videos of rainy scenes left. Most of the videos last around 3 *minutes*, and the rest last 90 *seconds*.

2.3 Selection by Two-stage Filtering

We adopt a two-stage filtering way to select target images that reflect the character of the rainy day scene as much as possible.

Stage i After finding out these raining videos, we first view each video and manually note down the approximate time intervals according to our understanding of rainy scenes. Then we extract full frames from the time intervals of clips we have recorded in the first step, containing 24,335 images in total.

Stage ii We apply BiSeNetV2 [Yu *et al.*, 2021] segmentation models to quickly filter all of these frames from stage i to acquire corresponding segmentation results. The purpose of this step is to allow the model to provide its perspective on these data to know what kind of situation the current models can not handle. During this stage, besides the droplets impact, we have discovered extra degradation factors, which help us split the dataset into three parts in the next step. The details of these factors will be outlined later. Lastly, we make the final selection, picking 500 images from stage i by referring to the segmentation results from stage ii. Among those 500 images, 300 mainly focus on diverse rainfall driving scenes, 100 of which comprise the unique windshield wiper blurring effect, and the rest 100 contain the windshield reflection situation.

2.4 Annotation Procedure

We annotate our images based on the label tool of Cityscapes, so the class split also follows its manner. Annotating an image on a heavily rainy day is difficult because of too many complex factors (*e.g.*, blurring and reflection). To make annotation more accurate and reliable, we propose using temporal information to enhance the quality: During the annotation, we will also refer to adjacent frames to decide the class of unclear pixels in current frame.

Following the construction of Cityscapes [Cordts *et al.*, 2016] and ACDC [Sakaridis *et al.*, 2021] datasets, we invite 20 annotators with professional image processing experience. In particular, annotated samples by one person are passed to another for a second confirmation to eliminate the bias and misinterpretation.

The time interval we choose in the selection part is usually at least 2 *seconds* long with 60 frames, which means that at least 60 adjacent continuous frames have accompanied every image we annotate and at least one video of reference. By combining spatial and temporal information, our annotation has richer information and reliable accuracy.

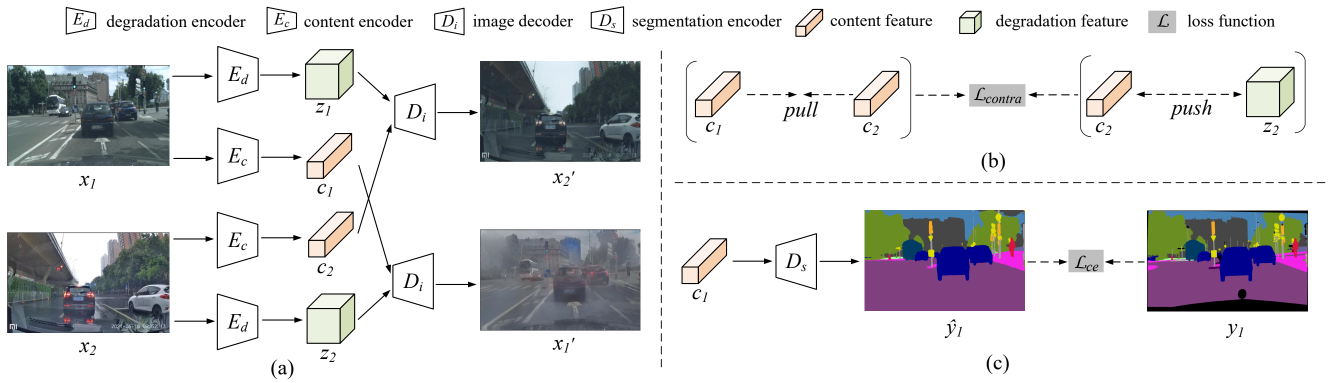


Figure 3: Overall of our framework. Our framework is divided into two steps. First, the rainy image, together with a clear scene image, is sent into the feature disentanglement network, where a degradation encoder encodes two specific degradation features and content encoder encodes two content features. After cross-reconstruction by exchanging the two degradation features, we achieve the basic function of deraining (a). Then, we view the (c_1, c_2) as the positive pair and (c_2, z_2) as the negative pair, proposing a contrastive loss $\mathcal{L}_{contra}$ as an extra constraint (b). Besides, the regular cross-entropy loss $\mathcal{L}_{ce}$ is used to ensure the initial performance of the segmentation model (c).

2.5 Rain Pattern Classification

To better study the scenes under specific rain intensity, we classified images. Specifically, we randomly selected 5,000 images and invited 20 relevant staff to rate each image with respect to rain intensity, according to the definition of Wikipedia, and finally, we assigned each image a specific intensity class based on the average score. This resulted in a total of three classifications: light rain, moderate rain, and heavy rain. We trained a classifier to classify the remaining images.

2.6 Comparison to Related Dataset

Finally, our dataset mainly contains 500 fine-grained pixel-level annotations, accompanied by corresponding 24,335 images of driving on rainy days, comprising diverse scenes and various degradation factors. The comparison of data diversity to other relevant driving datasets is shown in Table 1. The mixture of multiple elements has brought the most close-to-reality rainy scenes to Rainy WCity.

3 Proposed Method

The architecture of our framework is brief and straightforward, as shown in Fig. 3. Simply rely on structure information may cause ambiguous mapping problems [Liao *et al.*, 2021]. So we need to extract features in a more specific way. Motivated by [Ma *et al.*, 2022], we design a feature disentanglement network as the basic unit of our method. Given images x_1 and x_2 from Cityscapes and our Rainy WCity domain, with the "shared content space" assumption [Huang *et al.*, 2018], it can disentangle domain-invariant content features c_1 and c_2 of these images from the domain-specific counterparts z_1 and z_2. As has been validated by [Chang *et al.*, 2019], the content features contribute most to the semantic segmentation task. Therefore, through feature disentanglement, we can transfer segmentation knowledge from x_1 domain to x_2 domain.

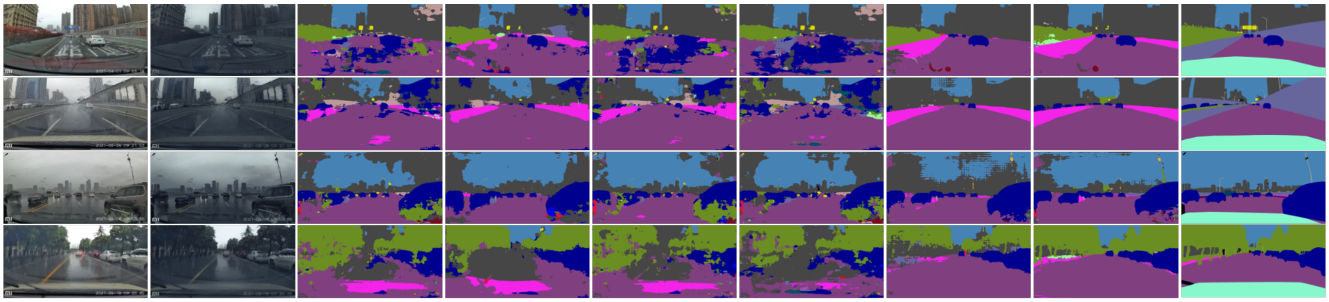
Specifically, we first need a shared content encoder E_c to extract c_1 and c_2 and two degradation encoders to extract degradation feature z_1 and z_2, respectively. Then, we use a shared image decoder D_i to decode an image using the content features c_1, c_2, and degradation feature z_1, z_2. Depending on which c and z we use, we can perform cross-reconstruction to supervise the disentanglement learning. Besides, we use a segmentation decoder D_s to produce segmentation heatmaps $\hat{y}_1$ from the content feature c_1, where label y_1 is used as the supervision signal. After the cross-reconstruction, we achieve the basic function of deraining of our framework.

The traditional way of deraining is to train a model to decompose the rain image into the clean background and residual of the rain streak. Since pair clean label is difficult to obtain, we integrate the task of deraining into a segmentation framework, as described in the first step. This idea is based on the assumption that the segmentation and deraining share a similar goal: deraining is to restore clean image, which segmentation can benefit from it. Consequently, the segmentation result of the derained image should outperform the original rain-degraded image. Specifically, we propose to utilize contrastive learning paradigm to distinguish the boundary between positive and negative pairs, where the comment latent content feature pair (c_1, c_2) as the positive pair and (c_2, z_2) as the negative pair. During training, we update the degradation encoders and segmentation encoders simultaneously. In this manner, the deraining module will study to restore rainy image progressively. Meanwhile, the segmentation network will also learn to discover the semantics in the restored images. The two modules complement each other in the process of training.

The goal of contrastive learning is to learn an encoder that encodes the private degradation pattern in our Rainy WCity dataset and helps disentangle the content, which contributes most to the semantics, from the rainy image. We assume that latent content feature c_1 and c_2 should be pulled as near as possible in the feature space, while content feature c_2 and degradation feature z_2 should be pushed as far as possible. Correspondingly, the definition of contrastive loss is as follows:

$$\mathcal{L}_{contra} = -\log \frac{\exp(c_2 \cdot c_1/\tau)}{\exp(c_2 \cdot c_1/\tau) + \exp(c_2 \cdot z_2/\tau)}, \qquad (1)$$

where $\cdot$ denotes the inner product, and $\tau > 0$ is a temperature hyper-parameter. Note that all the features in the loss function are $\ell 2$-normalized.



(a) Input (b) Derain (ours) (c) V3+ (d) V3+ (derain) (e) DualGCN (f) MPRNet (g) Our (derain) (h) Ours (i) Ground Truth

Figure 4: Qualitative comparison with DeepLabV3+ (V3+ for short). (c) is the result of DeepLabV3+ on our Rainy WCity dataset. For (d) to (g), restoration procedures are directly applied to the rainy images to generate corresponding rain-free outputs, and then we apply the publicly available pre-trained models of DeepLabV3+ for the segmentation task.

Method	PSPNet	DeepLabV3+	SDCNet	HRNet	BiSeNetV2
Rainy Input	13.46	23.21	**23.07**	26.89	25.93
DualGCN	12.75	23.01	22.56	26.81	25.82
MPRNet	**14.98**	21.83	21.99	**26.92**	23.98
S2R2 w deraining	11.87	25.88	19.58	25.31	27.82
S2R2 w trained model	-	**39.67**	-	-	**32.79**

Table 2: Comparison results of different segmentation models in terms of mIoU (%) on our Rain WCity dataset. Five segmentation models are directly applied to the original rainy inputs and the derained images produced by the DualGCN and MPRNet methods to perform the two-stage solutions for the rainy segmentation task. In particular, we also investigate combinations of our proposed unsupervised deraining model and these five segmentation methods to construct multiple one-stage rainy segmentation baselines.

This loss function can make the output of the deraining network closer to the ground truth and increase the distance between the restored image and the rain image in semantic space, which improves the classification result of scene pixels on the other hand.

In addition, we also introduce a pixel-level cross-entropy loss $\mathcal{L}_{ce}$ to constrain the semantic segmentation task. This loss will check each pixel one by one and compare the prediction result $\hat{y}_1$ of each pixel category with our label y_1. The definition of loss is as follows:

$$\mathcal{L}_{ce} = -\sum_{h,w}\sum_{c\in C}(\hat{y}_1^{(h,w,c)}\log y_1^{(h,w,c)}), \tag{2}$$

where y_1 are ground truths annotations for images x_1, $\hat{y}_1 \in \mathbb{R}^{H\times W\times C}$ is the segmentation softmax output of x_1, and H, W, and C represents height, width, and number of class categories, respectively.

The final overall loss function is:

$$\mathcal{L}_{final} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{contra}, \tag{3}$$

where λ is the weighting parameter.

4 Experimental Results

To investigate the segmentation task under rainy days and evaluate our proposed S2R2 model, we conduct extensive comparison experiments with existing representative methods on our proposed Rainy WCity dataset. In detail, it consists of two deraining methods (DualGCN [Fu *et al.*, 2021] and MPRNet [Zamir *et al.*, 2021]) and five segmentation models (HRNet [Wang *et al.*, 2021], BiSeNetV2 [Yu *et al.*, 2021], DeepLabV3+ [Chen *et al.*, 2018], SDCNet [Zhu *et al.*, 2019], and PSPNet [Zhao *et al.*, 2017]). For the segmentation part, all these comparison models are pre-trained on Cityscapes [Cordts *et al.*, 2016] and run by their default settings.

4.1 Evaluation Metric

We use mean Intersection over Union (mIoU) to evaluate segmentation effects; The higher the value, the better. Since it is almost impossible to get clean pairs for real rainy images, we adopt two no-referenced image quality assessment methods for evaluation of the effectiveness of derain models: Spatial–Spectral Entropy-based Quality (SSEQ) and Natural Image Quality Evaluator (NIQE), the smaller the value, the better for these two assessments.

4.2 Comparison Results on Rainy WCity

The rainy segmentation task involves two sub-tasks: image deraining and scene segmentation. To investigate the intrinsic contribution of these two tasks to the final segmentation performance, we construct several two-stage competitive models by directly cascading different deraining models (DualGCN, MPRNet, and our deraining model) and segmentation algorithms (HRNet, BiSeNetV2, DeepLabV3+, SDCNet, and PSPNet). In particular, we further design a one-stage method based on our deraining model and DeepLabV3+/BiSeNetV2 to promote the compatibility of these two sub-tasks via joint optimization. Quantitative evaluation on our proposed Rainy WCity dataset is tabulated in Table 2.

We use five segmentation models as baseline models, all of which fail to generate satisfied segmentation results. We guess the image deraining may destroy the intrinsic semantic information since these two sub-tasks have the obvious optimization discrepancy. We then test two deraining methods (DualGCN and MPRNet) and the deraining part (trained degradation encoder and image decoder) of our proposed method to see whether *deraining the rainy images as a pre-processing is helpful for the segmentation*. However, in contrast to the baseline models where the rainy images are directly input, the deraining processing has unstable effects, *i.e.*, some have improved the performance while others lowered the performance, proving the pre-processing of deraining is not an effective way.

Category	Method	road	sidew.	build	wall	fence	pole	light	sign	veget.	terrain	sky	person	rider	car	truck	bus	motorc.	bicycle	mIoU
Droplet	DeepLabV3+	73.5	7.5	47.4	7.4	11.5	24.6	6.9	46.5	47.0	0	78.1	5.1	0.2	43.6	1.5	0	0.1	1.5	21.2
	MPRNet	58.6	1.9	44.2	**12.4**	10.5	**25.7**	6.0	44.7	45.9	0	84.8	4.3	0.5	31.2	1.5	0.2	0.8	2.4	19.8
	DualGCN	73.8	**8.1**	50.5	6.3	7.3	23.3	6.9	46.6	46.3	0	82.3	5.4	0.7	41.8	1.0	0.5	0.1	1.3	21.1
	S2R2 (Ours)	**91.8**	6.9	**58.3**	3.2	**38.8**	18.1	**39.2**	**52.3**	**75.2**	**0.5**	**88.8**	**33.2**	**4.0**	**81.8**	**24.2**	**17.0**	**14.4**	**32.1**	**37.7**
Wiper	DeepLabV3+	63.8	0.5	47.4	6.3	0	21.9	0	25.8	42.8	0	86.6	0	0	28.9	0	3.8	0	0	17.2
	MPRNet	54.1	0	47.3	**9.6**	0	20.9	0	25.7	40.7	0	86.4	0	0	24.3	0	3.4	0	0	16.4
	DualGCN	66.3	**2.6**	48.7	3.4	0	15.8	0	28.8	40.8	0	85.5	0	0	32.4	0	5.0	0	0	17.3
	S2R2 (Ours)	**94.9**	**2.6**	**58.7**	4.7	0	**27.6**	0	**58.3**	**75.6**	0	**92.9**	0	0	**86.9**	0	**40.0**	**50.0**	0	**32.9**
Reflection	DeepLabV3+	73.5	4.8	46.2	9.8	26.0	20.0	14.5	38.7	57.0	0	86.0	7.9	0	45.7	1.1	6.0	0	0	23.0
	MPRNet	67.1	3.3	44.9	**13.2**	22.9	**20.7**	17.7	35.2	56.2	0	87.4	5.9	0	39.9	1.1	3.5	0.1	0	22.1
	DualGCN	72.2	4.6	45.5	9.0	11.2	20.2	14.6	**39.3**	54.1	0	88.6	9.2	0	43.5	1.9	4.0	0	0	22.0
	S2R2 (Ours)	**87.1**	**11.5**	**60.4**	9.0	**60.2**	20.1	**36.0**	22.2	**79.8**	0	**91.6**	**28.1**	**6.1**	**81.5**	**77.3**	**8.3**	**1.9**	0	**37.8**

Table 3: mIoU comparison is performed separately for each of the three categories of our dataset using DeepLabV3+. The first row of each category means implementing the model to our dataset. For the second and third rows, we apply two state-of-the-art derain methods to obtain restored images on our dataset, respectively, and then apply DeepLabV3+ on these images. The last row shows our one-stage method. Our proposed solution gains the best scores in almost all classes.

Category	Method	SSEQ ↓	NIQE ↓	mIoU ↑
Droplet	Original	29.5	4.1	21.2
	MPRNet [Zamir *et al.*, 2021]	29.5	4.1	19.8
	DualGCN [Fu *et al.*, 2021]	20.7	3.7	21.1
	S2R2 (Ours)	**20.5**	**3.4**	**37.7**
Wiper	Original	32.3	4.7	17.2
	MPRNet [Zamir *et al.*, 2021]	32.3	4.7	16.4
	DualGCN [Fu *et al.*, 2021]	**22.8**	4.2	17.3
	S2R2 (Ours)	25.3	**3.8**	**32.9**
Reflection	Original	28.1	4.0	23.0
	MPRNet [Zamir *et al.*, 2021]	28.1	4.0	22.1
	DualGCN [Fu *et al.*, 2021]	20.4	3.7	22.0
	S2R2 (Ours)	**20.1**	**3.2**	**37.8**
All	Original	21.9	3.8	23.2
	MPRNet [Zamir *et al.*, 2021]	29.7	4.2	21.8
	DualGCN [Fu *et al.*, 2021]	**21.0**	3.8	23.0
	S2R2 (Ours)	21.1	**3.4**	**39.7**

Table 4: SSEQ and NIQE comparisons between different derain methods on our dataset. The lower, the better. "Original" means unprocessed rainy images in Rainy WCity. Images with lower SSEQ and NIQE values are more likely to get good segmentation results.

The segmentation part (trained content encoder and segmentation decoder) of our proposed unsupervised method brings about significant improvement, which shows the effectiveness of our method.

Visual comparison are provided in Fig. 4. As expected, our one-stage segmentation model shows significant superiority over other two-stage competitors, producing the predicted maps with clearer objective boundaries and cleaner contents. Especially for the wiper and reflection scenarios, only our proposed model can achieve the accurate segmentation of the road and cars, while other comparison methods gain the dirty segmentation results.

To further analyze the individual influence on each class, we tabulate the mIoU in Table 3, involving a total of 19 classes commonly. It is evident that our proposed one-stage solution gains the best scores in almost all classes. Especially for the autopilot related classes, *e.g.*, road, building, car, and truck, our proposed method achieves significant improvements over other two-stage methods on the real rainy segmentation task. Taking the road, car, and bus as examples, the comparisons of the mIoU between ours and the second-best method are 73.8% (DualGCN) *vs* 91.8% (ours), 43.6% (DeepLabV3+) *vs* 81.8% (ours), and 0.5% (DualGCN) *vs* 17.0% (ours), respectively. These results reveal two key points: 1) simply cascading the deraining and segmentation models is far from producing the satisfying segmentation results on the real rainy days, our proposed Rainy WCity dataset in particular; 2) our proposed unsupervised one-stage scheme can significantly alleviate the compatibility, including the scenes (synthetic and real rainy conditions) and tasks (deraining and segmentation tasks), extensively promoting the final segmentation performance under real rainy days.

Table 4 provides the joint evaluation of image deraining and segmentation performance. Our proposed one-stage method achieves impressive performance on almost all metrics, gaining the best scores of the NIQE and mIoU in all scenarios. Although the two-stage method adopts DualGCN to promote the image quality and gains the smallest values of SSEQ, the final segmentation performance is undesired, even worse than that of the original input. We guess that the deraining operation may destroy the spatial semantic information, which is momentous for the segmentation task. By contrast, we adopt the joint optimization of the image deraining and segmentation, which greatly promotes the compatibility by achieving significant improvement in terms of the final mIoU (39.9% *vs* 23.2%). The metrics of deraining are negatively correlated with the metric of segmentation, which verifies our design.

5 Conclusion

This paper presents Rainy WCity, a semantic segmentation dataset for car-driving under the diverse real rainfall scenes. Apart from the common rainy situation, we have illustrated its complexity by proposing the fine-grained annotation of the driving circumstance where droplets, windshield reflection, and the windshield wiper exist, which are ignored in previous datasets but play a significant impact factor in this weather.

Besides an unsupervised joint optimization framework S2R2 is aimed to improve segmentation performance. Extensive experiments demonstrate the effectiveness of our method in real rainy scenarios.

Acknowledgments

This work was supported in part by the Department of Science and Technology, Hubei Provincial People's Government under Grant 2021CFB513 and 2021CFB281, and in part by the National Natural Science Foundation of China under Grant 62171325.

References

[Alhaija *et al.*, 2018] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars M. Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *Int. J. Comput. Vis.*, 126(9):961–972, 2018.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020.

[Chang *et al.*, 2019] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *CVPR*, pages 1900–1909, 2019.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, 2018.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Fu *et al.*, 2021] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *AAAI*, pages 1352–1360, 2021.

[Hu *et al.*, 2021] Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Trans. Image Process.*, 30:1759–1770, 2021.

[Huang *et al.*, 2018] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[Huang *et al.*, 2020] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2702–2719, 2020.

[Jiang *et al.*, 2020] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, pages 8343–8352, 2020.

[Jin *et al.*, 2021] Jiongchao Jin, Arezou Fatemi, Wallace P. Lira, Fenggen Yu, Biao Leng, Rui Ma, Ali Mahdavi-Amiri, and Hao (Richard) Zhang. RaidaR: A rich annotated image dataset of rainy street scenes. *arXiv abs/2104.04606*, 2021.

[Li *et al.*, 2019] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K. Tokuda, Roberto Hirata Junior, Roberto Marcondes Cesar Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, pages 3838–3847, 2019.

[Liao *et al.*, 2021] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Image inpainting guided by coherence priors of semantics and textures. In *CVPR*, pages 6539–6548, 2021.

[Ma *et al.*, 2022] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *CVPR*, 2022.

[Porav *et al.*, 2020] Horia Porav, Valentina-Nicoleta Musat, Tom Bruls, and Paul Newman. Rainy screens: Collecting rainy datasets, indoors. *arXiv abs/2003.04742*, 2020.

[Sakaridis *et al.*, 2021] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pages 10745–10755, 2021.

[Tung *et al.*, 2017] Frederick Tung, Jianhui Chen, Lili Meng, and James J. Little. The raincouver scene parsing benchmark for self-driving in adverse weather and at night. *IEEE Robotics Autom. Lett.*, 2(4):2188–2193, 2017.

[Wang *et al.*, 2021] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2021.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2633–2642, 2020.

[Yu *et al.*, 2021] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.*, 129(11):3051–3068, 2021.

[Zamir *et al.*, 2021] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 6230–6239, 2017.

[Zhu *et al.*, 2019] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn D. Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *CVPR*, pages 8856–8865, 2019.