The Raincouver Scene Parsing Benchmark for Self-Driving in Adverse Weather and at Night

Frederick Tung[1], Jianhui Chen[1], Lili Meng[2], and James J. Little[1]

***Abstract*—Self-driving vehicles have the potential to transform the way we travel. Their development is at a pivotal point, as a growing number of industrial and academic research organizations are bringing these technologies into controlled but real-world settings. An essential capability of a self-driving vehicle is environment understanding: where are the pedestrians, the other vehicles, and the drivable space? In computer and robot vision, the task of identifying semantic categories at a per pixel level is known as scene parsing or semantic segmentation. While much progress has been made in scene parsing in recent years, current datasets for training and benchmarking scene parsing algorithms focus on nominal driving conditions: fair weather and mostly daytime lighting. To complement the standard benchmarks, we introduce the Raincouver scene parsing benchmark, which to our knowledge is the first scene parsing benchmark to focus on challenging rainy driving conditions, during the day, at dusk, and at night. Our dataset comprises half an hour of driving video captured on the roads of Vancouver, Canada, and 326 frames with hand-annotated pixelwise semantic labels.**

***Index Terms*—Object detection, segmentation, categorization; Semantic scene understanding; Performance evaluation and benchmarking**

I. Introduction

Self-driving vehicles have the potential to transform the way we travel and even the way we design our cities. They are expected to cut traffic accidents, reduce congestion, enable better resource utilization, and increase the mobility of the elderly and disabled.

The safe automated operation of a vehicle requires robust environment sensing and understanding. This includes an accurate analysis of the semantic structure of the environment using visual sensors. Given an input image or video stream, the task of identifying and labelling semantic categories of interest, at a per pixel level, is known as *scene parsing* or *semantic segmentation* [1], [2], [3], [4], [5], [6], [7], [8], [9]. Scene parsing differs from the related task of object detection [10], [11] in two main ways. First, in object detection the typical output of the algorithm is a set of bounding boxes around detected objects, while in scene parsing the output is a dense pixelwise labelling of the scene. An accurate dense pixelwise labelling requires precise estimation of the spatial boundaries of semantic categories in the scene, which are valuable for such tasks as determining drivable free-space, navigating structured environments, and interacting with objects [12]. Second, object detection generally focuses on foreground objects with well-defined spatial extent (e.g. people, vehicles), while scene parsing is concerned with labelling both foreground objects and amorphous background elements (e.g. pavement, vegetation).

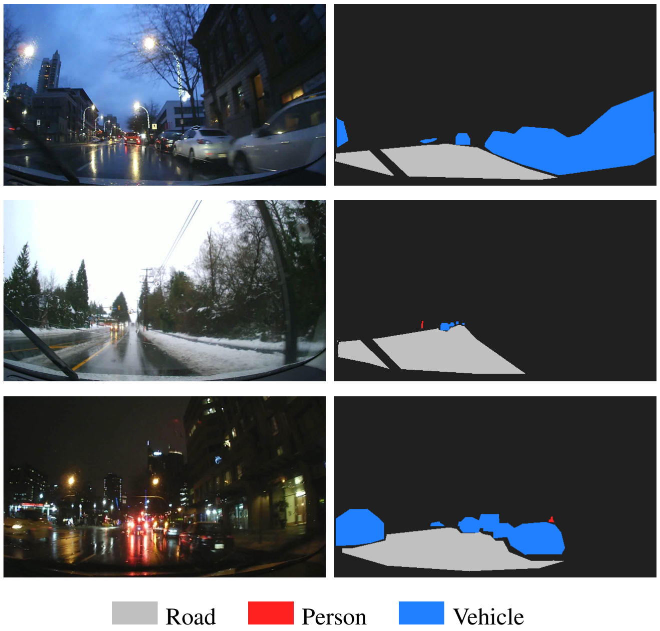

Fig. 1. Sample frames and annotations from the Raincouver scene parsing benchmark. We collect driving sequences in the rain and across a range of illumination conditions, including during the day, at dusk, and at night. To our knowledge, Raincouver is the first scene parsing dataset to focus on robust labelling in adverse weather and illumination conditions.

In the past few years, much progress has been made in scene parsing for self-driving. However, current systems are limited in application to fair weather and daytime illumination conditions. The standard scene parsing benchmarks for self-driving [13], [14], [15] are similarly focused on these nominal driving conditions. Robust operation in nominal driving conditions is an essential first milestone. Looking ahead, we believe that the next important challenge for self-driving is robust operation in adverse weather conditions, during the day, at dusk, and at night.

With this goal in mind, we have collected, manually annotated, and make publicly available a new scene parsing dataset

Manuscript received: Feb. 13, 2017; Revised May 10, 2017; Accepted June 22, 2017.

This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments. This work was funded in part by the Natural Sciences and Engineering Research Council of Canada.

[1] F. Tung, J. Chen, and J. J. Little are with the Department of Computer Science, University of British Columbia, Vancouver, Canada. {ftung, jhchen14, little}@cs.ubc.ca

[2] L. Meng is with the Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada. lilimeng@mech.ubc.ca

Digital Object Identifier (DOI): see top of this page.

for self-driving that specializes in rainy weather conditions. To our knowledge, our Raincouver benchmark is the first scene parsing benchmark to address the challenge of robust labelling in adverse weather and illumination. Raincouver includes sequences captured during the day, at dusk, and at night. As a specialized and smaller benchmark, Raincouver is designed to complement the standard scene parsing benchmarks for self-driving, by providing challenging sequences captured under weather and illumination conditions not covered in the conventional benchmarks. Fig. 1 shows some sample frames from three different video sequences in the benchmark.

The rest of this paper is organized as follows: we next review the task of scene parsing and describe the most commonly used scene parsing benchmarks for self-driving; we then discuss the collection of Raincouver and its ground truth annotation; we explain the composition of the benchmark and the technical challenges posed by Raincouver that are not present in the standard benchmarks; we report reference results on the benchmark using a fine-tuned deep neural network [5] and a nonparametric label transfer technique [4]; and finally, we present ideas for future work.

II. Related Work

Scene parsing, or semantic segmentation, refers to the task of identifying and localizing semantic categories of interest in an image or video, at a per pixel level [1], [2], [3], [4], [5], [6], [7], [8], [9] (though beyond the scope of this work, scene parsing can also be applied to point clouds and meshes [16], [17]).

The CamVid benchmark [13] was the first video-based dataset for scene parsing. CamVid contains over 700 densely annotated images covering ten minutes of driving video. Video sequences are collected using a digital camera mounted on the dashboard of a car.

The KITTI vision benchmark [14] is a broad ranged dataset for self-driving research, with a focus on stereo, optical flow, visual odometry, and 3D object detection. Multiple sensor modalities are provided, including laser range finder and GPS. Driving sequences are captured in Karlsruhe, Germany. Although KITTI does not include a standard set of dense pixelwise annotations for scene parsing, several research groups have independently labelled subsets of the dataset. For instance, Valentin et al. [16] and Sengupta et al. [18] labelled 70 images with the semantic categories of road, building, vehicle, pedestrian, pavement, vegetation, sky, signage, pole, and wall/fence, for the task of 3D scene parsing [9], [19].

The Cityscapes benchmark [15] is a very recently introduced large-scale dataset for self-driving, and consists of 25,000 densely annotated images of scenes in 50 cities – an extensive amount of data for training state-of-the-art deep models. By design, Cityscapes focuses on daytime and good weather conditions. The authors suggest that adverse weather and illumination conditions "require specialized techniques and datasets" [15].

We have collected and annotated Raincouver as a specialized dataset to fill this gap, with a focus on self-driving in adverse weather, including at dusk and at night. Our benchmark includes 326 hand annotated images covering half an hour of driving video, making it comparable in scope to CamVid. To the best of our knowledge, Raincouver contains more images than any publicly released labelled subset of KITTI.

Previous work in the autonomous navigation community has explored methods to achieve robustness in severe weather and dynamic illumination conditions, but in the context of localization (place recognition) and mapping. For example, Milford and Wyeth [20] developed a simultaneous localization and mapping method that achieves robust place recognition by matching sequences of frames. Sünderhauf et al. [21] introduced the Nordland dataset, comprising video sequences captured across four seasons from a railroad train, for simultaneous localization and mapping. Linegar et al. [22] proposed an efficient "life-long" learning technique for robot localization that supports changes in weather, illumination, and scene structure over time. Maddern et al. [23] recently introduced the Oxford RobotCar dataset, capturing 1000 km of repeated traversals of a fixed route in a wide range of weather and illumination conditions, as well as structural changes over time. These methods contribute important datasets for localization and mapping, however they do not include dense pixelwise labels.

Pfeiffer et al. [24] explored the problem of robust stereo vision and introduced a stixels dataset that includes adverse weather and illumination.

III. Collection and Composition Details

Our dataset consists of video sequences captured by a dashboard camera mounted behind the windshield of a 2014 Toyota Corolla. Images are captured at 720p resolution. The dashboard camera natively records at 30 frames per second and encodes the video using Motion JPEG. Since this produces enormous files, we have resampled the videos at 10 frames per second following the KITTI visual odometry dataset [14], and encoded the videos using MPEG-4 at a bitrate of 2000 kbps.

We construct ground truth annotations in the form of dense pixelwise labels, once every 60 frames (or 6 seconds), starting from the first frame in each sequence. We choose to annotate all images in-house by hand to ensure labelling quality and consistency. While a broad range of semantic categories may be useful in self-driving, we concentrate our in-house labelling effort to three categories that are essential for any self-driving system: vehicles, people, and roads. The remaining pixels in an image are marked unlabelled or void, and can be considered as part of the background (or part of the car). In total, 85.3% of the pixels in the dataset are unlabelled, 10.7% are labelled road, 3.9% are labelled vehicle, and 0.1% are labelled person.

Table I shows a summary of the four video sequences forming the training (or fine-tuning) set and the six video sequences forming the testing set. Specializing in adverse weather and illumination conditions, Raincouver is comparatively small in terms of raw images. We anticipate and suggest that data-intensive deep learning based methods may pre-train on a larger general-purpose dataset (e.g. Cityscapes [15]) and then use our training set for fine-tuning. We allocate the same

TABLE I
SUMMARY OF RAINCOUVER BENCHMARK SEQUENCES

Seq. ID	Duration	# Labelled Frames	Weather	Illumination
Training/Fine-Tuning Set				
Train-1	4:00	41	Rain	Day
Train-2	4:00	41	Clear	Dusk
Train-3	4:00	41	Rain	Dusk
Train-4	4:00	41	Rain	Night
Total	*16:00*	*164*		
Testing Set				
Test-1	2:40	27	Rain	Day
Test-2	2:40	27	Rain	Day
Test-3	2:40	27	Rain	Dusk
Test-4	2:40	27	Rain	Dusk
Test-5	2:40	27	Rain	Night
Test-6	2:40	27	Rain	Night
Total	*16:00*	*162*		

amount of driving time to the training (fine-tuning) set and the testing set, and roughly the same number of images with ground truth annotation. The training sequences cover rainy weather conditions under a full range of lighting conditions: day, dusk, and night. To test in more varied locations and scenarios, we make the testing sequences shorter and include more testing sequences. All driving sequences are captured in Vancouver, Canada, but there is no overlap in the locations of the ten sequences.

The weather and illumination conditions in our driving sequences pose some new technical challenges that are not present in conventional self-driving scene parsing benchmarks:

- *Limited visibility.* Rain and late time of day lead to severely decreased visibility in many of the driving scenes. The resulting deterioration in visual appearance cues makes the recognition task more difficult than in conventional benchmarks. At night, it can be particularly difficult to detect people in the absence of a strong contrast to surroundings (e.g. a person in front of a backlit bus stop), or motion cues.
- *Windshield wipers.* Similar to CamVid our camera is mounted on the dashboard behind the windshield. As a result, the camera picks up the motion of the windshield wipers in nine of the ten video sequences. The wipers may occlude important parts of the scene from view. Moreover, the frequency of their activation may change as the rain intensity changes. For simplicity, we mark the wiper pixels as unlabelled or void in the ground truth annotation.
- *Severe glare.* Many of our video sequences are captured at dusk and at night, when the vision system is challenged by severe glare from streetlights, traffic lights, and vehicle headlights. Oncoming vehicles with activated headlights often appear as a bright set of lights. We label them as vehicles in the ground truth nonetheless. The scene parsing algorithm has to learn to make use of visual context in order to distinguish oncoming vehicles from other sources of light and glare.
- *Partial snow cover.* In two of our video sequences, light snow cover can be seen on the road and on parked vehicles, making them more difficult to recognize. Robust scene parsing in snow covered urban environments is an important research direction, especially for the mass adoption of self-driving vehicles in colder climates, and is not yet addressed in the standard benchmarks. We do not explore this direction further in Raincouver but we believe a benchmark specializing in self-driving in the snow would be interesting for future work.

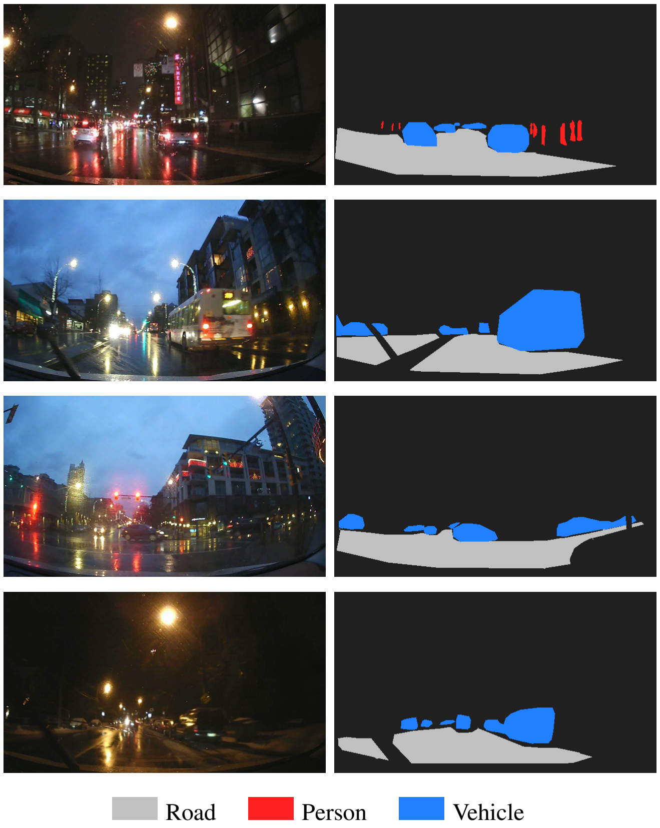

Fig. 2. Sample frames illustrating technical challenges present in the Raincouver video sequences, including limited visibility due to rain or time of day; occlusion from windshield wipers; severe glare at dusk and at night; and partial snow cover on roads and vehicles.

Fig. 2 shows some examples of the challenges described above.

IV. Baseline Experiments

We next provide two reference scene parsing baseline results for experimental comparisons. For a modern deep learning baseline, we fine-tune SegNet [5] on our training images with pre-training on either Cityscapes [15] or Synthia [25], a large synthetic dataset. The labels in Cityscapes and Synthia are converted to the Raincouver labels (vehicles, people, roads, and unlabelled/void) for pre-training the network. SegNet [5] is a fully convolutional neural network architecture designed for 2D scene parsing, comprising an encoder network, a corresponding decoder network, and a pixelwise classification layer. The encoder network follows the VGG-16 network architecture [26]. The decoder performs upsampling based on the max-pooling indices computed during encoding, and further convolves with learned filters.

CollageParsing [4] is a nonparametric scene parsing algorithm that formulates scene parsing from a search perspective instead of classification. Region proposal windows in the test image are matched with the most similar region proposal windows in the labelled database images, and the pixelwise labels of the matched database windows are "transferred" to the test image. More precisely, this label transfer is used to update the unary potentials in a conditional random field, which is solved to obtain the final image labelling. The search-based formulation allows CollageParsing to be easily extended to new database images and semantic categories without the need for re-training. We implement the CollageParsing algorithm with slight modifications, updating the region proposal generation and conditional random field inference to use more recent techniques [10], [27].

We adopt recall and intersection-over-union measures [16], [18] to evaluate scene parsing accuracy for each of the three categories. For a particular category c, recall is computed by counting the true positive pixelwise labels (true positives are pixels labelled as c in the ground truth and as c by the algorithm) and dividing by the sum of true positives and false negatives (false negatives are pixels labelled as c in the ground truth but differently by the algorithm). Intersection-over-union is a stricter measure that penalizes false positives. For a particular category c, the intersection-over-union is computed by counting the true positives and dividing by the sum of true positives, false negatives, and false positives (false positives are pixels labelled as c by the algorithm but differently in the ground truth). We consider pixels that are labelled as some foreground category by the algorithm, but that are void or background in the ground truth, to be false positives as well.

Since the task of scene parsing involves estimating the spatial boundaries of semantic categories, a natural question may be whether a direct measure of boundary precision could be evaluated. However, since boundaries ignore category identities, such an evaluation is not straightforward.

Table II summarizes the scene parsing results on the testing images. Overall, error rates on our benchmark are quite high. Segmenting the people in the scenes is particularly challenging for both baseline algorithms, and becomes more difficult at low levels of illumination. It is possible that the training set contains too few person instances to effectively fine-tune the network (we identified and annotated 73 person instances in the training/fine-tuning set). Moreover, unless they are close to the camera, people occupy relatively few pixels in the image, and at low illumination they are especially hard to distinguish from the background. Fig. 3 shows several frames from the dusk and night test sequences illustrating the difficulty of the task. Quantitatively, person segmentation performance for both baselines is considerably worse at dusk and at night than during the day.

TABLE II
BASELINE [4], [5] RESULTS ON RAINCOUVER

	Road	Person	Vehicle
Recall			
Rain, Day (Test-1 and Test-2)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	89.2%	37.4%	75.3%
b. Pre-training on Synthia	92.7%	4.2%	51.3%
CollageParsing	92.0%	52.8%	79.9%
Rain, Dusk (Test-3 and Test-4)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	85.2%	4.9%	77.1%
b. Pre-training on Synthia	76.5%	1.9%	80.8%
CollageParsing	90.5%	29.1%	73.2%
Rain, Night (Test-5 and Test-6)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	82.2%	8.4%	80.7%
b. Pre-training on Synthia	72.9%	7.7%	79.3%
CollageParsing	91.7%	6.9%	73.7%
Overall			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	85.6%	17.2%	**78.2**%
b. Pre-training on Synthia	81.0%	4.9%	72.1%
CollageParsing	**91.4%**	**28.5**%	75.3%
Intersection-over-union			
Rain, Day (Test-1 and Test-2)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	75.8%	17.2%	38.3%
b. Pre-training on Synthia	69.4%	1.3%	24.4%
CollageParsing	71.9%	24.1%	46.6%
Rain, Dusk (Test-3 and Test-4)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	71.6%	2.9%	37.4%
b. Pre-training on Synthia	65.2%	0.9%	27.1%
CollageParsing	70.3%	7.0%	42.2%
Rain, Night (Test-5 and Test-6)			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	67.3%	6.0%	44.3%
b. Pre-training on Synthia	63.0%	2.7%	31.5%
CollageParsing	71.5%	5.4%	42.8%
Overall			
Fine-tuned SegNet			
a. Pre-training on Cityscapes	**71.6%**	9.9%	40.0%
b. Pre-training on Synthia	66.1%	1.8%	28.4%
CollageParsing	71.2%	**12.0%**	**43.7%**

However, we anticipate that better person detection may be possible by taking advantage of motion cues. Both of our baseline algorithms parse single frames in isolation and do not make use of temporal information, whether in the form of temporal prediction smoothing, motion clustering, or spatiotemporal appearance-based features [28], [29], [30]. Temporal information may be particularly useful in adverse driving conditions in which environment visibility is limited and distinctive visual appearance cues are intermittent. In fact, to produce high-quality ground truth annotations, our in-house manual annotation process involves re-playing the video around the target frame to minimize the chances of omitting low-contrast people (and vehicles, to a lesser extent). We find that, for our driving sequences, the task of identifying all person instances from only single frame information can be difficult even for a human annotator. While temporally consistent scene parsing is beyond the scope of this work, the complete video sequences are included in our dataset to enable potential methods to leverage information across multiple video frames.

We also performed experiments with SegNet without fine-tuning on our training images; however, the performance

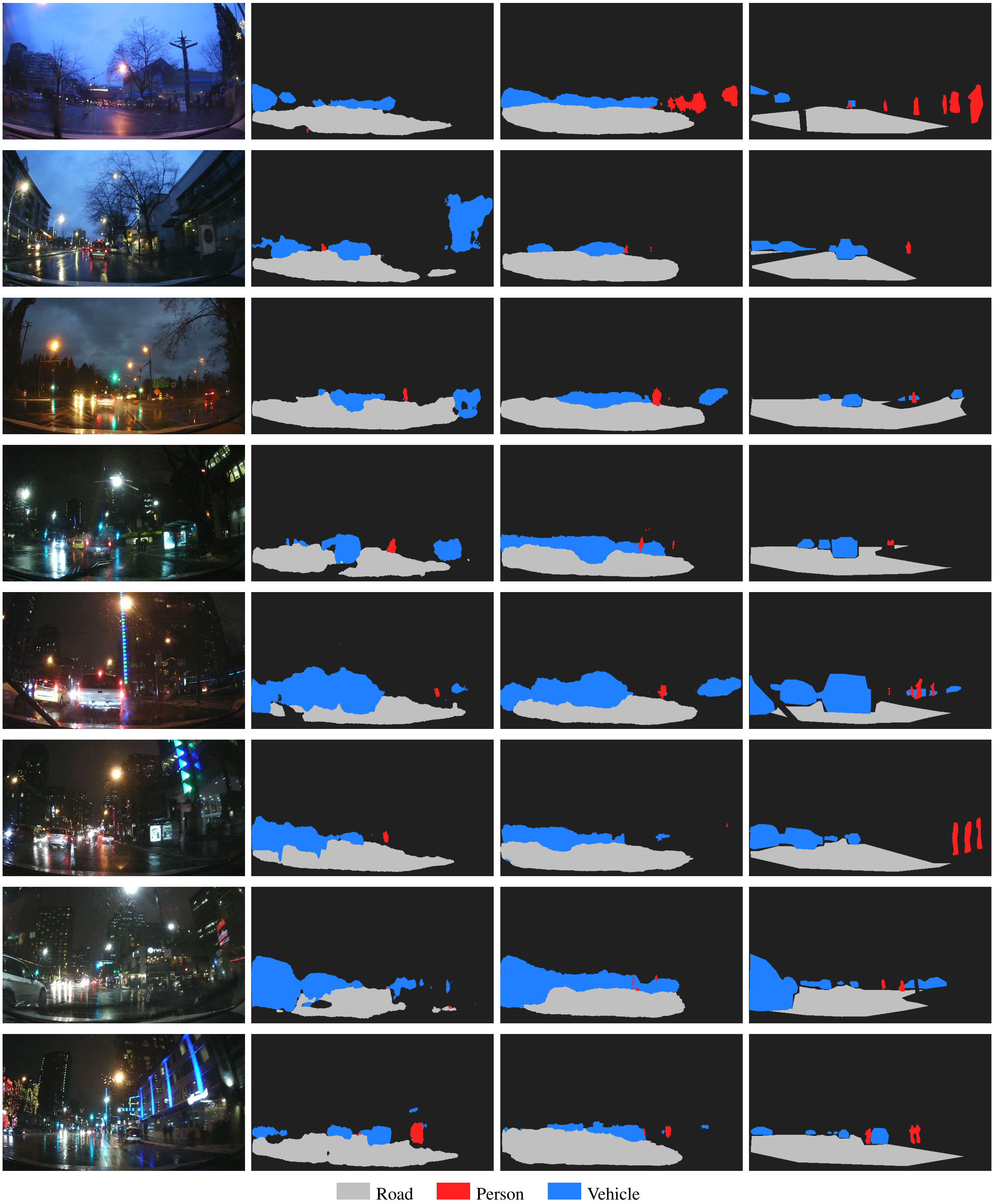

Fig. 3. Sample frames from the Test-3, Test-4, and Test-5 sequences illustrating the difficulty of person detection in rain at dusk and at night. From left to right: RGB image; SegNet [5] baseline output (fine-tuned with pre-training on Cityscapes); CollageParsing [4] baseline output; ground truth annotation.

is worse. Cityscapes-pretrained SegNet without fine-tuning achieves only 37.4%, 2.2%, and 8.8% intersection-over-union for road, person, and vehicle, respectively, compared to 71.6%, 9.9%, and 40.0% with fine-tuning. Synthia-pretrained SegNet without fine-tuning achieves 14.7%, 0.7%, and 8.8% intersection-over-union for road, person, and vehicle, respectively, compared to 66.1%, 1.8%, and 28.4% with fine-tuning. The drop in performance when fine-tuning is omitted suggests that training on conventional datasets is insufficient for scene parsing in adverse weather and illumination conditions. Running Cityscapes-pretrained SegNet on the Cityscapes validation set (converted to Raincouver labels) produces 88.9%, 53.4%, and 79.1% intersection-over-union for road, person, and vehicle, respectively – much higher accuracy than in Raincouver's adverse driving conditions. Scene parsing in the rain, and at dusk or at night, poses generalization challenges that motivate the use of a complementary dataset specializing in these conditions.

V. Conclusion and Future Work

We have collected and annotated the Raincouver scene parsing benchmark with the intention of supporting research efforts towards self-driving vision systems that are robust to adverse weather and illumination conditions. Our specialized benchmark is complementary to the standard scene parsing benchmarks for self-driving, and we advocate its use in combination with them to cover a broader and more realistic range of challenging driving conditions.

Safe and reliable operation in adverse conditions is necessary for the mass market adoption of self-driving vehicles. In practice, a self-driving vehicle will be equipped with additional sensors, such as LIDAR, IMU, and GPS, which we have not considered here. We believe that an important direction for future work will be the fusion of different sensor modalities for more robust scene understanding in adverse weather and illumination conditions.

References

[1] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011.

[2] J. Yang, B. Price, S. Cohen, and M. Yang, "Context driven scene parsing with attention to rare classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] M. Najafi, S. T. Namin, M. Salzmann, and L. Petersson, "Sample and filter: nonparametric scene parsing via efficient filtering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] F. Tung and J. J. Little, "Scene parsing by nonparametric label transfer of content-adaptive windows," *Computer Vision and Image Understanding*, vol. 143, pp. 191–200, 2016.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," arXiv:1511.00561, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional neural networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision*, 2015.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *International Conference on Learning Representations*, 2015.

[9] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *IEEE International Conference on Robotics and Automation*, 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] S. Sengupta, P. Sturgess, L. Ladický, and P. H. S. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[13] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. S. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[17] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[18] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *IEEE International Conference on Robotics and Automation*, 2013.

[19] F. Tung and J. J. Little, "MF3D: Model-free 3D semantic scene parsing," in *IEEE International Conference on Robotics and Automation*, 2017.

[20] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation*, 2012.

[21] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons," in *Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation*, 2013.

[22] C. Linegar, W. Churchill, and P. Newman, "Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation," in *IEEE International Conference on Robotics and Automation*, 2015.

[23] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000km: The Oxford RobotCar dataset," *International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[24] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[25] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[27] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2011.

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[29] O. Miksik, D. Munoz, J. A. Bagnell, and M. Hebert, "Efficient temporal consistency for streaming video scene analysis," in *IEEE International Conference on Robotics and Automation*, 2012.

[30] A. Bewley, V. Guizilini, F. Ramos, and B. Upcroft, "Online self-supervised multi-instance segmentation of dynamic objects," in *IEEE International Conference on Robotics and Automation*, 2014.