Soiling detection for Advanced Driver Assistance Systems

Filip Beránek[1][0009−0002−3563−3097], Václav Diviš[1][0000−0001−9935−7824], and Ivan Gruber[1][0000−0003−2333−433X]

Department of Cybernetics and New Technologies for the Information Society, Technická 8, 301 00 Plzeň, Czech Republic bionothi@ntis.zcu.cz

Abstract. Soiling detection for automotive cameras is a crucial part of advanced driver assistance systems to make them more robust to external conditions like weather, dust, etc. In this paper, we regard the soiling detection as a semantic segmentation problem. We provide a comprehensive comparison of popular segmentation methods and show their superiority in performance while comparing them to tile-level classification approaches. Moreover, we present an extensive analysis of the Woodscape dataset showing that the original dataset contains a data-leakage and imprecise annotations. To address these problems, we create a new data subset, which, despite being much smaller, provides enough information for the segmentation method to reach comparable results in a much shorter time. All our codes and dataset splits are available at https://github.com/filipberanek/woodscape_revision.

Keywords: Semantic segmentation, Automotive, Data-Centric AI, Soiling detection

Introduction

n pursuit of safer roadways, automotive manufacturers are using the power of Adanced Driver Assistance Systems (ADAS) [31]. These systems, ranging from Adaptive ruise Control (ACC) to Lane-Keeping Assist (LKA), represent a critical step toward lly autonomous vehicles [10]. Additionally, comfort-oriented features like parking asistance further improve the driving experience. The central role of the functionality of DAS is playing various sensors such as lidars, radars, ultrasound, and cameras. While hese sensors perform reliably under optimal weather conditions, real-world scenarios resent challenges. Factors like occlusion, particularly when sensors are exposed and unshielded, can significantly compromise system reliability.

Until recently, there has been a small number of research addressing poor visibility conditions. This changed with the introduction of the UG2 competition at CVPR 2019 [27]. Among others, the Valeo company addressed this research gap by releasing the first dataset featuring occluded camera imagery, known as the Woodscape dataset [28]. Alongside the dataset they released extensive work on occlusion detection [24][7][25][26].

Soiling detection extends beyond mere identification. It encompasses evaluating the level of occlusion and location. Consider a scenario where a small portion of an image

is heavily obscured, impacting crucial decisions like lane-keeping or collision avoidance. Hence, understanding the position and level of soiling becomes imperative.

Building upon Valeo's pioneering work, we extend their research. Previous articles [24][7] mention semantic segmentation as a potential solution, but only tile-level detection has been explored. In this paper, we expand on the previous research by conducting baseline experiments with semantic segmentation models. Moreover, we provide an indirect comparison to published results by Valeo.

Lastly, we dive into the Woodscape dataset. Starting with the dataset and its split revision, where we would like to explore how the data split is done. Then we explore the impact of annotation quality. Given the subjective nature of soiling annotation, where the same occlusion could be identified by different annotators as different levels of occlusion, achieving precise distinctions between classes poses a challenge, as discussed in [25] and [26]. We aim to evaluate annotation quality, present basic dataset statistics, and assess how variations in annotation quality affect semantic segmentation performance.

Our contributions can be summarized as follows:

- A comprehensive analysis of the Valeo's Woodscape dataset, including dataset split refinement.
- An introduction of semantic segmentation results and their comparison to tile-level detection.
- An examination of the influence of annotations refinement on the segmentation accuracy.

2 Related Work

Research related to occlusion or adverse conditions has three main directions. Firstly, denoising strategies endeavor to eliminate occlusions from camera images. For instance, Porav et al. [19] focus on removing water droplets from camera lenses. Other notable works include those by [11][6][2][14][23], all dedicated to restoring images to their original clarity by mitigating occlusions.

Secondly, efforts are made to adapt networks in order to work better under adverse conditions. Works such as those by Sakaridis et al. [22][21][18][1] focus on foggy scenes, striving to execute semantic segmentation accurately under challenging conditions.

Thirdly, the detection of occlusions, primarily spearheaded by Valeo, is highlighted in the following works [24][7][25][26].

All these approaches face a common obstacle: the scarcity of data. This situation has been alleviated by initiatives like [27]. Additionally, Valeo Woodscape dataset [28], the first soiling dataset published, has paved the way for further research in this domain.

Most of the previous research focused predominantly on developing new and improving already existing algorithms. However, the most important aspect of supervised learning, the data, was neglected. Introduced by Andrew Ng [17], Domain or Data-Centric AI (DCAI) is currently increasing in importance in AI research. There has been also published Six principles of DCAI by Jarrahi et al. [13] and further evaluation of large-scale Automotive datasets, conducted by Divis et al. [8]. This research highlighted the potential influence of unbalanced features and proposed methods how to evaluate and collect uniformly feature-distributed domain-related datasets.

3 Dataset analysis

The Woodscape dataset initially comprised 76,448 images and their corresponding labels [28][24], however, only a fraction of this dataset has been made publicly available, totaling 5,000 images with the labels. The dataset is divided into two subsets: the "Train set," containing 4,000 images, and the "Test set", which contains 1,000 images. Notably, the images were captured sequentially, typically featuring 7 or 8 consecutive frames from the same scene and camera setup. In the dataset, there are four following classes used in annotations: *Clear*, *Transparent*, *Semi-Transparent*, *Opaque*. The class distribution is heavily imbalanced, where each class contains approximately the following number of pixels: *Clear* - 39% of pixels, *Transparent* - 12%, *Semi-Transparent* - 19%, and *Opaque* - 30%.

3.1 Data split

Even though the original dataset is split into test and train subsets, we don't recommend using the same structure. We identify a data leakage between these two original folders. As was already mentioned, images are taken in sequence, where usually 7 or 8 images are from the same scene in sequence and from the same camera. These 7 or 8 images are split between the train and test folder, which means we would be training the network on the same scene as we are later concluding the test.

As there is a data leak between the Valeo train and test subset, we provided our own data split, which addresses this problem. We collected all images and made split based on sequences. We also reduced the number of test images, so we have a bigger training set totalling 4503 images for training and 497 images for test. It should be noted that when we are referring to the train and test set in the rest of the paper, we are always referring to this new refined split.

3.2 Class definition revision

Within the original dataset, we encounter four distinct labels as mentioned in [26] with the following definitions:

- *Clear*: indicates no soiling present.
- *Semi-transparent* and *Transparent*: are blurry regions where background colors are visible with diminished texture. The difference between the *Semi-transparent* and *Transparent* classes is the ability to recognize the object through the blur.
- *Opaque*: represents that in that particular region, it is impossible to see through.

However, the differentiation between the *Transparent* and *Semi-transparent* classes is somewhat ambiguous according to this definition.

To address this ambiguity, we propose a refined classification scheme, related to the soiling:

- *Clear*: Represents scenarios without any kind of occlusion.
- *Transparent*: Encompasses regions where colors may be altered by occlusion or objects may appear blurred, yet lines and structural elements maintain their integrity. Additionally, colors remain consistent but may exhibit tonal variations. Examples could be a small deposit of dust.

– *Semi-transparent*: Describes regions where object deformities are evident in contrast to *Transparent* regions. For example, distorted lines or unrecognizable colors may be present, though objects remain visible and distinguishable. Example could be more dust on camera or water drops.
– *Opaque*: Signifies regions where visibility is entirely obscured by non-transparent colors, preventing observation of what lies beyond. Example could be water drop that are distorting section of image that much, as its not possible to see through. Another example could be mud or leafs.

An illustrative example of these annotations can be seen in Fig. 1

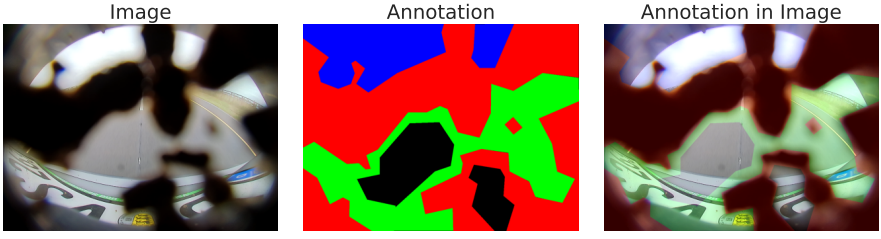

Fig. 1: Example of a soiling annotation from the Woodscape dataset. Black is *Clear*, green is *Transparent*, blue is *Semi-transparent*, and red is *Opaque* class

3.3 Annotation revision

As aforementioned the Woodscape dataset was not flawlessly annotated. The following study [26] points out the potential mismatches in annotations. Additionally, articles such as [24] and [7] address inconsistencies and challenges in distinguishing between classes. Consequently, we had to conduct a manual inspection of the dataset, identifying three key labeling scenarios:

– Image is correctly annotated.
– Completely erroneous annotations, that could be confusing. An example of such annotations is illustrated in Fig. 2, depicting a cut-off annotation in the first row, affecting the following 685 images.
– Ambiguities in annotations; which are especially notable within the *Transparent* and *Semi-transparent* classes, where the distinction between the two classes is not sharp. The lack of clear demarcation between these classes can potentially impact the outcomes. Such an ambiguity affected additional 519 cases.
– Inconsistencies in class definitions, where the same scene may be labeled as *Transparent* in one instance and *Semi-transparent* in another. This inconsistency, often driven by subjectivity, contributed to further training and evaluation challenges. Approximately 1375 images were affected by such inconsistencies.

Examples of these annotation challenges are depicted in the Fig. 2.

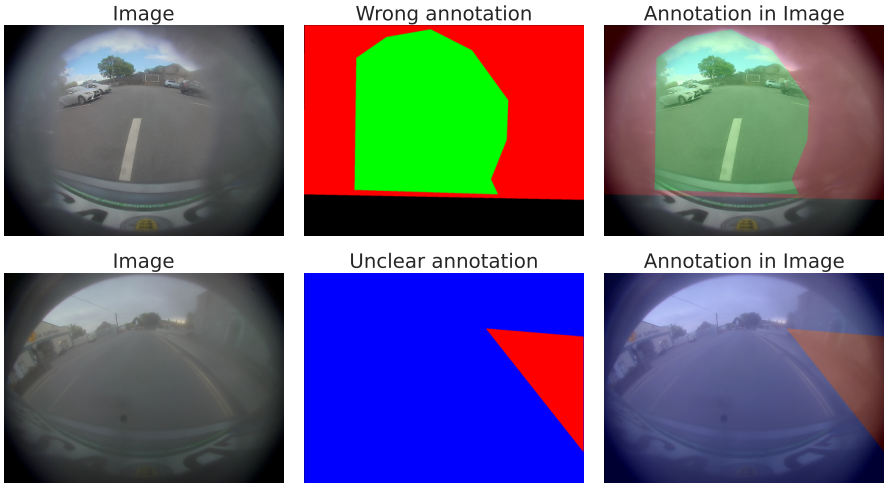

Fig. 2: Example illustrating the issues with annotations. The first row exhibits a clear annotation error, while the second row showcases a simplification in class differentiation.

To address this problem, we provide three additional training subsets based on the precision of the annotations in them. We denote our original split from Subsection 3.1 as *Baseline set*. After that, we take the training part of the *Baseline set* and split it into two parts - training set and validation set. Each of these subsets is then manually cleaned by an expert annotator. We opt to let only one expert clean the whole dataset so we prevent subjective differences between two different experts. At the end of this procedure, we get four training and four validation sets in total. The distribution of class-related pixels for these sets can be found in Table 1.

Table 1: Split of data into train and validation sets, and distribution of classes for each set. Table is sum of class related pixels in millions.

Annotation class	Training set	Validation set
Clear	1981	378
Transparent	716	36
Semi-transparent	1133	53
Opaque	1702	142

The presented subsets are defined as follows:

- *Baseline set*- comprising all files, including those with errors.
- *Correct files* - excluding obvious errors.

- *Correct clear files* - eliminating unclear annotations and simplifications in addition to errors.
- *Correct clear strict files* - further refining by removing inconsistencies in class assignments.

It should be noted that we provide only one original test set for all the subsets. This step ensures a fair comparison between these subsets. The class distribution of the individual subsets can be found in Table 2.

Table 2: Split of data into train and val sets, and distribution of classes for each set. Table is sum of class related pixels in millions.

Baseline set			Correct files		
Class	Training	Validation	Class	Training	Validation
Clear	1773	208	Clear	1649	177
Transparent	590	125	Transparent	463	87
Semi-transparent	926	206	Semi-transparent	675	181
Opaque	1419	282	Opaque	1079	222
Correct clear set			Correct clear strict files		
Class	Training	Validation	Class	Training	Validation
Clear	1500	177	Clear	889	130
Transparent	281	69	Transparent	121	21
Semi-transparent	523	134	Semi-transparent	154	13
Opaque	924	145	Opaque	375	62

4 Experiments

In this chapter, we present a series of experiments. Firstly, we focus on the task of semantic segmentation enriching the previous results from Valeo. Within these experiments, we also study the impact of different training setups, namely the choice of encoder size and choice of loss function. Secondly, we validate the training data and its impact on the segmentation results. As mentioned in Section 3, we provide four subsets of data addressing the problem with annotations strictness. We compare the performance of the chosen semantic segmentation method and also its training t

If not stated otherwise, we are using the following experimental setup. We start with a learning rate of 0.01 and schedule it with a patience factor of 5 epochs with ϵ of 0.2. We also use an early stop condition with a patience of 10 epochs and a minimum loss bigger than 0. Learning rate scheduler as well as early stop conditions are related to validation accuracy. The maximum number of epochs is set to 400. We use Resnet18 [12] as an encoder. Batch size is equal to 5. As an optimizer, we use Adam with ϵ 1^{-8}. Finally, we conduct training on CPU AMD Ryzen 7 7840HS having RAM 32GB LPDDR5x and graphic card NVIDIA GeForce RTX 4060 with 8GB VRAM.

4.1 Networks experiments

Two notable approaches and their outcomes, as documented by Valeo in [24] and [7] can be used as a baseline for our further experiments. SoilingNet utilizes a ResNet10 encoder and a unique decoder designed for tile-level classification. This setup employs tile dimensions of 64 x 64 pixels. Both models were trained and tested on the complete dataset comprising 76,448 images. However, discrepancies arise in the label definitions used, as only three labels—"Clean," "Transparent," and "Opaque"—are employed, differing from the annotations in the published dataset and subsequent works such as [7]. Moreover, the 64 × 64 pixels tile size employed by SoilingNet poses challenges for comparison with semantic segmentation methodologies.

In contrast, TiledSoilingNet [7] adopts the same label definitions but operates at a higher classification granularity, namely a tile size of 4 x 4 pixels. This finer granularity aligns better with semantic segmentation, allowing us to compare the results. It should be noted that the results are not directly comparable, because Valeo used the whole datasets in their previous research, whereas in our research we are only able to use its small publicly-available fraction.

To be able to compare our approaches with the proposed baseline, we employ an accuracy metric despite not being optimal for unbalanced datasets. All our listed results are the results from the test set.

In this study, we utilize the following popular semantic segmentation approaches: Unet [20], Unet++ [30], DeepLabV3 [4], DeepLabV3++ [5], FPNet [16], PSPNet [29], MaNet [9], LinkNet [3], and PAN [15].

As can be seen in Table 3 the best performance was achieved by FPNet with an accuracy of 94.0%. All the tested semantic segmentation methods surpass the original tile-level classification approach.

Table 3: Accuracy of TileSoilingNet providing tiled classification based on 4x4 pixel matrix compared to semantic segmentation networks using same loss and similar size of encoder as aforementioned.

Architecture	↑ Accuracy
TiledSoilingNet[7]	0.874
Deeplabv3[5]	0.912
Unet++[30]	0.925
Deeplabv3+[4]	0.927
Unet[20]	0.929
Pan[15]	0.930
MANet[9]	0.934
PSPNet[29]	0.934
LinkNet[3]	0.936
FPNet[16]	**0.940**

To further explore possibilities of semantic segmentation approaches we take the best approach - FPNet - and test it with encoders of five different sizes. The results can be found in Table 4. It can seen that the difference between encoders is mostly

negligible. We believe that this phenomenon is caused by the relatively small size of the training dataset. That means that bigger encoders can't utilize their better computational capacity.

Table 4: Experiment displaying the correlation between encoder size and results on Woodscape dataset using FPNet.

Encoder	↑Accuracy
Resnet18	0.940
Resnet34	0.942
Resnet50	**0.945**
Resnet101	0.935
Resnet152	0.939

As mentioned in [7], another factor that can affect the result is the loss function. In the original paper, the authors claim that the best loss for this task was RMSE. To verify this conclusion, we conducted an additional ablation study using different losses as depicted in Table 5. In our experimental setup, the best results are reached while training with Cross-Entropy loss, nevertheless, the differences between different setups are not statistically significant again.

Table 5: Experiment displaying correlation of loss and accuracy using FPNet with different loss functions.

Loss function	↑Accuracy
Focal	0.939
Dice	0.930
Cross-Entropy	**0.944**
RMSE	0.940

4.2 Training sets experiments

In this subsection, we focus on different data subsets. As mentioned earlier, we create four different data subsets based on the annotation's strictness. We choose the five best networks from the previous subsection to compare their performance further. We train all of the networks with a Categorical Cross-entropy loss function. The rest of the training hyperparameters remain unchanged.

A comparison of training accuracies and training times per one epoch can be found in Fig. 3 and Fig. 4, respectively.

The performance for different subsets is comparable except the last subset - *Correct clear strict files* where the significant drop in performance can be observed for the most tested methods. We argue this is expected behavior due to the large difference between

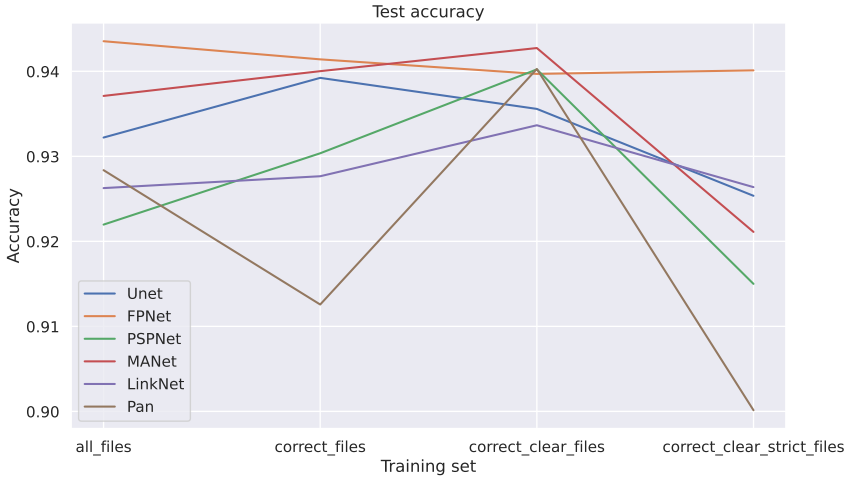

Fig. 3: Measured test accuracies across all networks and different training sets.

Fig. 4: Measured training time per epoch across all networks and different training sets.

the number of files in the last subsets. For the best trade-off between the training time and the model's accuracy, the usage of *Correct clear files* subset is recommended.

5 Conclusion

In this paper, we have presented a deep analysis of the Woodscape dataset that reveals several issues. The first issue - data leakage - was addressed by proposing a new data split. An additional analysis revealed vague class definitions, which can cause trouble while annotating new data. We extended these original class definitions to make them more precise and, therefore, prevent possible problems in the future. Lastly, we revealed problems with already existing annotations, and based on the manual inspection of the data we created four new subsets of the training data.

In the second part of the paper, we experimented with chosen semantic segmentation methods on the problem of soiling detection. These experiments show the superiority of the semantic segmentation approach while all of the tested methods surpass the previous research, which handles the soiling detection as tile-level classification. Moreover, we provide an ablation study on the topic of the choice of encoder size and also the choice of the loss function.

In our future research, we would like to focus on additional ablation studies of training hyperparameters and also test transformer-based approaches. Also obtaining additional training data can be beneficial for the whole scientific community.

Acknowledgements

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017 and by ARRK Engineering GmbH. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Alshammari, N., Akçay, S., Breckon, T.P.: Multi-task learning for automotive foggy scene understanding via domain adaptation to an illumination-invariant representation (2019)
2. Berman, D., Treibitz, T., Avidan, S.: Non-local image dehazing (06 2016). https://doi.org/10.1109/CVPR.2016.185
3. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE (Dec 2017). https://doi.org/10.1109/vcip.2017.8305148, http://dx.doi.org/10.1109/VCIP.2017.8305148
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation (2018)
6. Chen, Z., He, Z., Lu, Z.M.: Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention (2023)

7. Das, A., Krizek, P., Sistu, G., Burger, F., Madasamy, S., Uricar, M., Kumar, V.R., Yogamani, S.: Tiledsoilingnet: Tile-level soiling detection on automotive surround-view cameras using coverage metric (2020)
8. Diviš, V., Schuster, T., Hrúz, M.: Domain-centric adas datasets. In: SafeAI@AAAI (2023), https://api.semanticscholar.org/CorpusID:259100725
9. Fan, T., Wang, G., Li, Y., Wang, H.: Ma-net: A multi-scale attention network for liver and tumor segmentation (2020). https://doi.org/10.1109/ACCESS.2020.3025372
10. Fathy, M., Ashraf, N., Ismail, O., Fouad, S., Shaheen, L., Hamdy, A.: Design and implementation of self-driving car (01 2020). https://doi.org/10.1016/j.procs.2020.07.026
11. Fattal, R.: Single image dehazing, acm trans. Graph., SIGGRAPH **27** (01 2008)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Jarrahi, M.H., Memariani, A., Guha, S.: The principles of data-centric ai (dcai). arXiv preprint arXiv:2211.14611 (2022)
14. Ki, S., Sim, H., Choi, J.S., Seo, S., Kim, S., Kim, M.: Fully end-to-end learning based conditional boundary equilibrium gan with receptive field sizes enlarged for single ultra-high resolution image dehazing (06 2018). https://doi.org/10.1109/CVPRW.2018.00126
15. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation (2018)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017)
17. Ng, A.: Deep Learning AI data-centric ai competition. https://https-deeplearning-ai.github.io/data-centric-comp/ (2021), accessed: 2022-08-16
18. Pfeuffer, A., Dietmayer, K.: Robust semantic segmentation in adverse weather conditions by means of sensor data fusion (2019)
19. Porav, H., Bruls, T., Newman, P.: I can see clearly now : Image restoration via de-raining (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
21. Sakaridis, C., Dai, D., Hecker, S., Gool, L.V.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding (2018)
22. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (Mar 2018). https://doi.org/10.1007/s11263-018-1072-8, http://dx.doi.org/10.1007/s11263-018-1072-8
23. Singh, M.: Image reconstruction and edge detection based upon neural approximation characteristics (01 2013). https://doi.org/10.12720/joig.1.1.12-16
24. Uricar, M., Krizek, P., Sistu, G., Yogamani, S.: Soilingnet: Soiling detection on automotive surround-view cameras (2019)
25. Uricar, M., Sistu, G., Rashed, H., Vobecky, A., Kumar, V.R., Krizek, P., Burger, F., Yogamani, S.: Let's get dirty: Gan based data augmentation for camera lens soiling detection in autonomous driving (2020)
26. Uricar, M., Sistu, G., Yahiaoui, L., Yogamani, S.: Ensemble-based semi-supervised learning to improve noisy soiling annotations in autonomous driving (2021)

27. Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W.J., Wang, Z., Zhang, T., Zhong, Q., Xie, D., Pu, S., Zheng, Y., Qu, Y., Xie, Y., Chen, L., Li, Z., Hong, C., Jiang, H., Yang, S., Liu, Y., Qu, X., Wan, P., Zheng, S., Zhong, M., Su, T., He, L., Guo, Y., Zhao, Y., Zhu, Z., Liang, J., Wang, J., Chen, T., Quan, Y., Xu, Y., Liu, B., Liu, X., Sun, Q., Lin, T., Li, X., Lu, F., Gu, L., Zhou, S., Cao, C., Zhang, S., Chi, C., Zhuang, C., Lei, Z., Li, S.Z., Wang, S., Liu, R., Yi, D., Zuo, Z., Chi, J., Wang, H., Wang, K., Liu, Y., Gao, X., Chen, Z., Guo, C., Li, Y., Zhong, H., Huang, J., Guo, H., Yang, J., Liao, W., Yang, J., Zhou, L., Feng, M., Qin, L.: Advancing image understanding in poor visibility environments: A collective benchmark study. IEEE Transactions on Image Processing **29**, 5737–5752 (2020). https://doi.org/10.1109/TIP.2020.2981922
28. Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricar, M., Milz, S., Simon, M., Amende, K., Witt, C., Rashed, H., Chennupati, S., Nayak, S., Mansoor, S., Perroton, X., Perez, P.: Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving (2021)
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network (2017)
30. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation (2018)
31. Ziebinski, A., Cupek, R., Grzechca, D., Chruszczyk, L.: Review of advanced driver assistance systems (adas) (11 2017). https://doi.org/10.1063/1.5012394