Let’s Get Dirty: GAN Based Data Augmentation for Camera Lens Soiling Detection in Autonomous Driving

Michal Určičář1, Ganesh Sistu2, Hazem Rashed2, Antonín Vobecký3,2, Varun Ravi Kumar2, Pavel Krňáček1, Fabian Bürger2 and Senthil Yogamani2

1 Independent Researcher 2 Valeo 3 CTU in Prague

uricar.michal@gmail.com antonin.vobecky@cvut.cz firstname.lastname@valeo.com

Abstract

Wide-angle fisheye cameras are commonly used in automated driving for parking and low-speed navigation tasks. Four of such cameras form a surround-view system that provides a complete and detailed view of the vehicle. These cameras are directly exposed to harsh environmental settings and can get soiled very easily by mud, dust, water, frost. Soiling on the camera lens can severely degrade the visual perception algorithms, and a camera cleaning system triggered by a soiling detection algorithm is increasingly being deployed. While adverse weather conditions, such as rain, are getting attention recently, there is only limited work on general soiling. The main reason is the difficulty in collecting a diverse dataset as it is a relatively rare event.

We propose a novel GAN based algorithm for generating unseen patterns of soiled images. Additionally, the proposed method automatically provides the corresponding soiling masks eliminating the manual annotation cost. Augmentation of the generated soiled images for training improves the accuracy of soiling tasks significantly by 18% demonstrating its usefulness. The manually annotated soiling dataset and the generated augmentation dataset will be made public. We demonstrate the generalization of our fisheye trained GAN model on the Cityscapes dataset. We provide an empirical evaluation of the degradation of the semantic segmentation algorithm with the soiled data.

1. Introduction

Level 5 autonomous driving ([10]) stands out as a challenging goal of a large part of the computer vision and machine learning community. Due to this problem’s difficulty, a combination of various sensors is necessary to build a safe and robust system. Even just considering cameras, a combination of a narrow field of view (FOV) long-range sensing and wide FOV short-range sensing is necessary. The latter is achieved by fisheye cameras, which provide complete near-field sensing, up to 10m around the vehicle. One use-case where it is necessary is the fish-bone parking [9], where ultrasonic sensors and narrow FOV cameras are not sufficient. Without exaggeration, these fisheye cameras are becoming de facto standard in low-speed maneuvering applications, like parking and traffic jams assist [16]. Enormous progress can be noticed in typical image processing tasks, such as semantic segmentation or object detection [23, 24, 25], which is mainly attributed to the prevailing success of deep learning. However, there are other less “popular” problems slowly getting into attention, which have to be solved as well for the ultimate goal of the full Level 5 autonomy.

One of these problems is the reliability of the sensory signal, which in the case of surround-view cameras means, inter alia, the ability to detect soiling on the camera lens. Failure to recognize severe weather conditions leading to a deterioration of the image quality to such a level that any further image processing is unreliable. Figure 2 shows

Figure 1: The example of a semi-transparent soiling in form of a water drop on the camera lens. The detection of the bus behind the water drop works still well, while the road segmentation (green) is highly degraded in the soiled region. In this scenario, a soiling detection algorithm is used to trigger a camera cleaning system which restores the lens hardware.
how the surround-view camera can get soiled and the corresponding image output, as well as an example of images taken during a heavy rain. It usually happens when the tire is splashing mud or water from the road or due to wind depositing dust on the lens. Figure 1 shows an example of the strong impact of a significant water drop on the camera lens for object detection and semantic segmentation tasks.

In this work, we focus on soiling caused by a variety of unwanted particles reposing on the camera lens. The source of these particles is mostly mud, dirt, water, or foam created by a detergent. Based on the state of aggregation, such soiling can be either static (e.g., highly viscous mud tends to dry up very quickly, so it does not change its position on the output image over time) or dynamic (mostly water, and foam). Because of that, the acquisition of suitable data for training machine learning models or merely testing the effect on existing classification models is quite tedious. Human annotation is very time demanding, costly, and not very reliable since the precise labeling of soiling on a single image can sometimes be very challenging, e.g., due to the unclear boundary. We want to emphasize that the soiling detection task is necessary for an autonomous driving system as it is used to trigger a camera cleaning system that restores the lens [29]. It is complementary to building segmentation or object detection models robust to soiling without an explicit soiling detection step. Our contributions include:

- A proposition of a baseline pipeline for an opaque soiling generation, based on CycleGAN [34] and soiling segmentation learned from weak labels.
- A novel DirtyGAN network, which is an end-to-end generalization of the baseline pipeline.
- Public release of an artificial soiling dataset as a companion to the recently published WoodScape Dataset [33], coined Dirty WoodScape Dataset, to encourage further research in this area.
- An empirical evaluation of the degradation of semantic segmentation algorithms on soiled images.

The rest of the paper is organized as follows. Section 2 covers the related work and contrasts the soiling scenarios with adverse weather conditions. In Section 3, we give a detailed description of the proposed algorithms. Section 4 provides an evaluation of the quality of the generated images and quantifies the improvement of the soiling detection algorithm by adding generated images to training. Also, it describes the empirical evaluation of the degradation of semantic segmentation in the presence of soiling. Finally, Section 5 concludes the paper.

2. Related Work

As the visual perception modules for autonomous driving are becoming more mature, there is much recent effort to make them more robust to adverse weather conditions. It can be seen by the popular CVPR workshop “Vision for all seasons”¹, which focuses on the performance of computer vision algorithms during adverse weather conditions for autonomous driving. However, there is very little work on the related but different lens soiling problem. The two problems are similar in how they degrade image quality and can severely affect visual perception performance. Yet there are substantial differences.

The first significant difference is that soiling on the lens can be removed by a camera cleaning system that either sprays water or uses a more sophisticated ultrasonic hardware [29]. Secondly, there is temporal consistency for soiling where mud or water droplets remain static typically or sometimes have low-frequency dynamics of moving water droplets in contrast to higher variability in adverse weather scenes. Thus this temporal structure can be exploited further for soiling scenarios. Finally, soiling can cause more severe degradation as opaque mud soiling can completely block the camera.

We focus on the generic soiling detection task in this paper. Even disregarding camera cleaning, soiling detection is still needed to increase the uncertainty of vision algorithms in the degraded areas. The task of soiling detection on camera lenses in autonomous driving is shortly described in [27]. The authors present a sort of a proof of concept idea on how Generative Adversarial Networks (GANs) [7] could...

¹https://vision4allseasons.net/
be applied for dealing with the insufficient data problem in terms of an advanced data augmentation. In the same paper, the authors also outline another potential usage of GANs in the autonomous driving area. A more formal introduction to the soiling detection and categorization is provided in [29].

The problem is formalized as a multilabel classification task and also discusses the applications of soiling detection, including the camera cleaning. Utičáč et al. [28] provided a desoiling dataset benchmark.

The rest of this section provides an overview of the commonly used GAN based image-to-image translation framework, which we use in the proposed method. Due to the limited work on soil, we also review the closely related area of adverse weather scenarios.

### 2.1. GAN based Image-to-Image Translation

In recent years, the task of artificial image generation is dominated by GANs [7], showing a great ability to synthesize realistic-looking images in [18]. The Image-to-Image translation is a part of graphics and computer vision that aims to learn a mapping between a source domain $X$ and a target domain $Y$ with the use of paired data. In [11], the authors present a method using GANs to tackle the problem of image-to-image translation using paired data. However, obtaining such paired data can be difficult and sometimes even impossible. Therefore, an unsupervised version, without any examples of corresponding pairs, is even more critical and challenging.

This problem is tackled by CycleGAN [34] with the use of two mappings $G: X \rightarrow Y$ and $F: Y \rightarrow X$. Since these mappings are highly under-constrained, they propose to use a cycle consistency loss to enforce $F(G(X)) \approx X$. Even though it is not emphasized that much in the CycleGAN paper, the authors in their implementation use also identity losses, $G(Y) \approx Y$ and $F(X) \approx X$, which improve the results significantly.

### 2.2. Rainy Scenes

Rainy scenes are slightly related to water soil,ation. Because this degradation is semi-transparent and some background information is still visible, it is common to use an image restoration algorithm that improves the image’s quality. The work of [15] provides a comprehensive analysis of this topic. The authors of [32] address the problem of rain removal from videos by a two-stage recurrent network. The rain-free image is estimated from the single rain frame at the first stage. This initial estimate serves as guidance along with previously recovered clean frames to help obtain a more accurate, clean frame at the second stage.

In [21], the authors propose a progressive recurrent de-raining network by repeatedly unfolding a shallow ResNet with a recurrent layer. In [30], a dataset of $\approx 29.5k$ rain/rain-free image pairs are constructed, and a SPatial Attentive Network (SPANet) is proposed to remove rain streaks in a local-to-global manner. Porav et al. [19] presented a method that improves the segmentation tasks on images affected by rain. They also introduced a dataset of clear-soiled image pairs, which is used to train a denoising generator that removes the effect of real water drops. Li et al. [14] proposed a two stage algorithm incorporating depth-guided GAN for heavy rain image restoration. Quan et al. [20] use double attention mechanism CNN for raindrop removal.

### 2.3. Dehazing

Another type of image quality degradation is caused by the presence of aerosols (e.g., mist, fog, fumes, dust, . . . ) in the environment surrounding the car. Due to the light scattering caused by these aerosol particles, the resulting image tends to have faint colors and looks hazy, which can inherently also impact the further image processing.

Fattal presents in [6] a method for single image dehazing, based on a refined image synthesis model and a depth estimation. Berman et al. [2], on the other hand, propose a solution, which is not based on local priors and builds on the assumption that a dehazed image can be approximated by a few hundred distinct colors which form tight clusters in the RGB color space. Ki et al. [12] propose fully end-to-end learning-based boundary equilibrium GANs to perform an ultra-high-resolution single image dehazing. Yan et al. [31] propose a semi-supervised learning using a mixture of real data without ground truth and synthetic data.

### 3. Artificial Soiling Generation

The task of single image soil annotation on the fish-eye cameras is quite tedious. We make use of polygonal annotation, which is a compromise of annotation speed and quality. However, even this kind of polygonal annotation is sometimes tough to interpret even by human annotators. It is particularly true for the soil boundary, which is usually very fuzzy.

However, an even bigger problem is how to obtain the soil data. In our setup, we apply a random pattern of soil on the camera lens using a particular soil generator. Then drivers ride the car for a while and repeat the process several times. It has many limitations: Firstly, it is very inconvenient to record data for all probable scenarios (e.g., driving through the city, rural area, highway). Secondly, it is not possible to measure the real impact of soil on the images, because we need a clean version of the same images for a fair comparison.

All these limitations motivate us to use the synthetic soil data. In the following sections, the proposed soil generation algorithms are described.
Figure 3: Several examples from the WoodScape Dataset. Top: RGB images from the fisheye camera. Bottom: corresponding human made annotations. White polygons represent soiling masks, while the background is marked by black color.

Figure 4: A few examples of the segmentation network results. In the left column are original soiled images, in the middle are the coarse annotations, and on the right column are the masks obtained by the segmentation network. The clean area is marked in dark gray, the opaque area is white, and the semi-transparent region light gray.

3.1. Soiling Generation Baseline Pipeline

The core of our baseline pipeline is formed by a CycleGAN [34] network, which we train to perform the image-to-image translation from clean images to their soiled counterparts. The main problem of the CycleGAN method is that it modifies the whole image. For our desired application, this can lead to undesired artifacts in the generated images. Besides that, the generated synthetic soiling patterns are often relatively realistic. Note that due to GPU memory requirements and time constraints, our CycleGAN training uses rescaled images (1/4 of both width and height).

Next, we train a soiling semantic segmentation network, $M$, using the weak polygonal annotation of soiling (see Figure 3 for several examples). Even though the annotation is quite coarse, the segmentation network $M$ is able to fit the soiling patterns more precisely. See Figure 4 for a few examples of $M$ outputs in comparison to the original annotations. We use WoodScape Dataset [33] for training the soiling segmentation network. Last but not least, we train a super-resolution network $U$, which we use to transform the GAN generated image to the original image resolution (i.e., up-scaling of 4× factor).

The idea of a baseline data generation algorithm is described in Figure 5. We take the generator transforming a clean image to the soiled image ($G_{CS}$) and apply it to the clean image $I$. It gives us an image with a random soiling pattern $I_s$. Next, we obtain the soiling mask $m$. This is achieved by applying the semantic segmentation network on the generated soiled image followed by a Gaussian smoothing filter $\gamma$: $m = \gamma(M(I_s))$. The resulting soiling mask $m$ is an image with values in range $[0, 1]$, where 0 means background, and 1 means soiling. The intermediate values can be understood as semi-transparent soiling. We apply the Gaussian smoothing filter because it mimics the physical nature of the soiling phenomenon where the edges of the soiling patterns are typically semi-transparent, due to photon scattering. Finally, the artificially soiled version of the original image $\hat{I}$ is a composition of the original image $I$ and the soiling pattern $I_s$ via the estimated mask $m$:

$$\hat{I} = (1 - U(m)) \cdot I + U(m) \cdot U(I_s).$$  \hspace{1cm} (1)
Figure 5: The soiling generation baseline pipeline. From left to right: an image into which we would like to paint a random soiling pattern; CycleGAN generated “soiled” version; blurred mask of the segmented soiling from the generated image; the resulting artificially “soiled” image obtained by convex combination of the original image and the generated soiling via the segmented mask.

Note, it is possible to use arbitrary images for the final composition, once we obtain $I_s$ and $m$. The mask $m$ obtained by the semantic segmentation network $M$ serves as an automatic annotation of the soiling in the generated image.

This simple pipeline has certain limitations. The biggest one is that it cannot be expected to work smoothly for soil types caused by water (e.g., raindrops, snow) in this specific formulation. One option for dealing with this issue is to follow the approach of [1], where the authors model the reflection patterns of the water drops using the whole image and apply filters and transformations consequently. The other option is to formulate a CycleGAN-like approach, which can cope with changing only those parts of the image that correspond to the soiling pattern and keep the rest unchanged. We formulate this approach in the following section.

3.2. DirtyGAN

The problem of applying CycleGAN in artificial soiling synthesis is twofold. Firstly, CycleGAN does not constrain the image generation to any specific regions and instead re-gen-erates the whole image, affecting all pixels. In the case of the artificial soiling generation, this is highly undesirable. For the investigation of the soiling impact on the further image processing, the background (i.e., regions of the image not affected by the soiling) must remain untouched. Secondly, the generation branch “clean” → “soiled” is ill-defined, as there is no visual clue for where the soiling should be produced. There are infinitely many patterns that could be created. Furthermore, there is also no control over the soiling pattern production process.

The first problem can be addressed, e.g., by Insta-GAN [17]. However, in such a case, the second problem becomes even a more significant issue. We decided to guide the pattern generation process via a Variational AutoEncoder (VAE) [5] and modify the CycleGAN algorithm so that it applies only on the masked regions of the source and target domain images. We coin the proposed network DirtyGAN.

We use the weak polygonal soiling annotations from the WoodScape Dataset for training the VAE. The main idea of using VAE for the soiling patterns generation is as follows. By using the encoder of the trained VAE, we can obtain the projection of an actual sample from the dataset to a lower-dimensional representation. If we select two samples $z_1$ and $z_2$ that are close on the soiling pattern manifold, we can obtain a novel sample $z$ by taking their convex combination.

$$z = \alpha z_1 + (1 - \alpha) z_2,$$  \hspace{1cm} (2)

where $\alpha \in [0, 1]$. Then, we take this intermediate representation $z$ and apply the trained decoder from VAE to reconstruct the corresponding soiling pattern. In Figure 6, we depict several examples of this intermediate soiling pattern reconstruction.

The benefit of using sampling from the learned VAE is that we could even use it to create animated masks, e.g., to mimic dynamic soiling effects, such as water drops in heavy rain or to be able to investigate the impact of dynamic soiling in general. After training the VAE, we limit CycleGAN to be applied only on the masked regions corresponding to the generated mask for the “clean” → “soiled” translation or the mask obtained by the soiling semantic segmentation mask $M$. We use a similar composition as in the baseline presented in Section 3.1. In Figure 7, we depict the whole DirtyGAN scheme.

3.3. Dirty Datasets

We have used the baseline pipeline to generate artificial soiling on our recently published WoodScape Dataset [33] with 10k images, which comes with semantic segmentation annotation. It makes it a suitable candidate for the soiling generation since we can merge the provided annotation with the soiling mask and measure the direct impact on classification models. Our generated data with updated annotation will be released as a WoodScape Dataset companion under the name of Dirty WoodScape Dataset².

In Figure 8, we show several generated examples together with their automatically obtained annotations. Our method of soiling generation is not limited to fiseye images

²https://github.com/valeoai/WoodScape
Figure 6: Variational AutoEncoder and the walk on the soiling manifold. The leftmost column depicts the original soiling pattern (the annotation for some particular soiled image). The next column is the reconstructed version by applying the whole VAE. The next 10 columns represent the transition (the walk on the soiling manifold) from the leftmost image to the rightmost one, which represents another example.

Figure 7: The DirtyGAN scheme. The original CycleGAN scheme is enhanced by VAE for novel soiling pattern generation in “clean” → “soiled” branch and by mask estimation in the opposite branch. Please see the supplementary material for a higher resolution version of this image.

only. Since we would like to support standard benchmarking as well, we also release a Dirty Cityscapes dataset. It is, as the name suggests, based on the Cityscapes [4] dataset.

4. Experimental Evaluation

4.1. Data Augmentation for Soiling Detection Task

A primary purpose of generating soiling is to mimic the real soiling scenarios. Firstly, we trained a soiling segmentation network on real soiling test data only (8k images). We tested this model performance on real soiling test data and achieved 73.95% accuracy. Then, we added generated soiling data (8k images) to our training catalog. As the generated soiling replicates the real soiling patterns, the network’s performance increased to 91.71% on real soiling test data, forming a 17.76% increase in accuracy without the need for costly annotations and real-time soiling scene captures (see the second and fourth row of Table 1).

We observe a significant reduction in accuracy when the network is trained on artificial images solely. It can be attributed to a limited ability to capture the entire data diversity to support the original soiled data. We also conduct a simple ablation study with standard data augmentation techniques (flipping, contrast changes) to match the size of training data with generated samples and observe a much lesser improvement in accuracy (third row), compared to the scenario with augmentation by generated synthetic samples.

The classifier used in the experiment uses ResNet50 [8] for an encoder and FCN8 [22] for a decoder. The binary cross-entropy was used as a loss function with the ADAM optimizer [15] for training with a learning rate of $1 \times 10^{-4}$. The image resolution was 640 × 480 pixels.

Table 1: Comparison of Soiling Segmentation model trained on generated and real soiled images. Accuracy is computed on a real test dataset with 2,000 images.

<table>
<thead>
<tr>
<th>Soiling Segmentation Model</th>
<th>Accuracy [%] on real test 2K dataset (mIoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained solely on generated images (8000 samples)</td>
<td>47.41</td>
</tr>
<tr>
<td>Trained solely on real images (8000 samples)</td>
<td>73.95</td>
</tr>
<tr>
<td>Trained on real &amp; data augmentation (16000 samples)</td>
<td>78.20</td>
</tr>
<tr>
<td>Trained on real &amp; generated images (16000 samples)</td>
<td><strong>91.71</strong></td>
</tr>
</tbody>
</table>

4.2. Artificial Soiling Quality

We performed a subjective visual study of the quality of the artificial soiling similarly as it is done in other GAN based image generation algorithms. We selected represen-
Figure 8: Examples of the generated images from the DirtyWoodScape Dataset, together with the generated annotations. Top: WoodScape Dataset RGB images with generated opaque soiling. Bottom: corresponding automatically generated annotations. White pixels represent the soiling, while black pixels represent the background.

We chose a random subset of 5% of generated samples for quantitative comparison due to the manual annotation cost. We use the mIoU metric to compare between artificially generated annotation and manual annotation. We obtain a high accuracy of 93.2%. It is to be noted that manual annotation can be subjected, especially at boundaries, and sometimes it is worse than artificial annotation based on visual inspection.

4.3. Degradation Effect on Semantic Segmentation

To demonstrate the impact of the soiling on the camera lens, we provide an empirical evaluation on semantic segmentation task. Pixel level classification is commonly used in autonomous driving applications. Our dataset consists of 10k pairs of images (clean and synthetically soiled using the proposed baseline framework). We split our data into training and testing by 80 : 20 ratio according to the sampling strategy described in [26]. We trained two DeepLabV3+ [3] models on the clean and soiled images, respectively. We evaluate the performance separately on clean and soiled test data. Table 2 summarizes the obtained results.

A segmentation model trained on clean images records 56.6% mIoU on clean test data and 34.8% on soiled data, a performance drop of 21.8% compared to clean images. This significant drop shows that soiling can cause severe degradation to a standard visual perception task in autonomous driving.

A model trained on the synthetic soiling data shows a limited degradation to 16.1%. However, training on the soiled images shows a 4% accuracy degradation on clean test data compared to the baseline when evaluated on clean images. Note, that model trained on the synthetic data perform reasonably well on the clean data with a few percentage drop in mIoU. This is expected as the data contain additional class, which is treated as background/unknown (void class). Thus, the trained model used less portion of data to
train the remaining classes.

Figure 9 depicts a qualitative evaluation of the soil-
ing impact on the segmentation task. The baseline model
trained on clean images 9a is evaluated on soiled images in
9c showing a high level of degradation due to the soil-
ing. Figure 9d shows the ground truth annotations, while 9f
and 9g illustrate results of model trained on the soiled images
while testing on the clean images in 9a and soiled images
in 9e. In a realistic scenario, annotations are not available
for the occluded region of soiled images. Using our GAN
generated dataset, we use annotations in the soiled area to
enable the model to interpolate segmentation classes in oc-
ccluded soiled parts. Figure 9f shows the capability of seg-
mentation networks to perform segmentation even behind
the soiled area. However, it is less reliable compared to the
clean baseline and sensitive to overfitting.

The same type of experiments were conducted using the
Cityscapes dataset [4]. The results are presented in Table 3,
which, as you can see, show a similar trend as in the Wood-
Scape experiment.

5. Conclusions

In this paper, we proposed two algorithms for the gen-
eration of soiling on images from surround view fisheye
 cameras. The first algorithm is a pipeline built from sev-
eral well-known blocks, such as CycleGAN [34], semantic
segmentation of the generated soiling, and image composi-
tion. The second algorithm is a novel DirtyGAN network,
which can generate similar results as the baseline pipeline
in an end-to-end fashion. The possibility to generate ran-
dom but realistic soiling patterns on camera images is an
integral component in examining the degradation of other
image processing methods. We provided an empirical eval-
uation of the performance degradation on several typical
classification tasks common for autonomous driving sce-
narios. We demonstrate that our soiling model trained on
fisheye images generalizes well on the Cityscapes dataset,
enabling us to create dirty versions of public datasets. Last
but not least, we release a public dataset as a companion
to the recently published WoodScape Dataset [33], coined
Dirty WoodScape Dataset, which can serve as a benchmark
for measuring the degradation of the off-the-shelf classifi-
cation algorithms.

Acknowledgements. Antonín Vobecký was supported by
the Grant Agency of the Czech Technical University in
Prague, grant No. SGS18/205/OHK3/3T/37. All authors
would like to thank Valeo for the support of publishing.
References


[4] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 6, 8


