RoadSaW: A Large-Scale Dataset for Camera-Based Road Surface and Wetness Estimation

Kai Cordes[1], Christoph Reinders[2], Paul Hindricks[1], Jonas Lammers[1],
Bodo Rosenhahn[2], and Hellward Broszio[1]
[1]VISCODA GmbH
[2]Institute for Information Processing, Leibniz University Hannover
https://roadsaw.viscoda.com

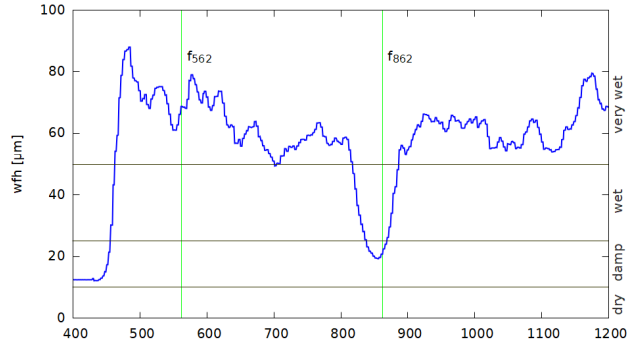
Abstract

Automated driving is one of the most promising technologies for improving road safety. In real driving scenarios, knowledge about the road friction is crucial. For the estimation of the road friction, two properties are of main interest: the road surface type and the road condition. We propose a novel large-scale dataset to enable camera-based road surface and wetness estimation. It consists of video data captured by in-vehicle cameras and ground truth for the current surface type and wetness which is determined by the MARWIS (Mobile Advanced Road Weather Information Sensor). The wetness measurements are associated to high-resolution bird's eye view road image patches, derived from a calibrated sensor setup. Additionally, data for different distances to the vehicle is provided. The dataset is evaluated with state-of-the-art real-time capable approaches for road condition classification and uncertainty estimation. The results provide a valid baseline, but also demonstrate limitations of the generalization performance. The dataset enables new possibilities for future research on camera-based road friction estimation. It is the first dataset including accurate measurements for the wetness in real driving scenarios.

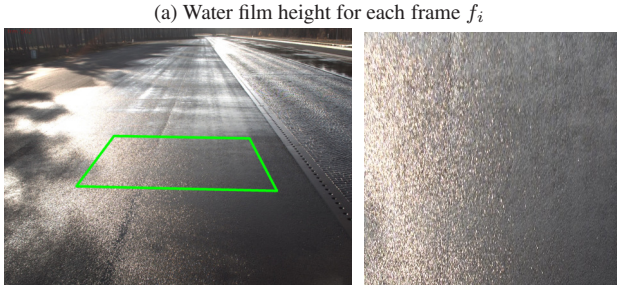
1. Introduction

Knowledge about the road friction is important for automated driving as well as for driver assistance systems to react reliably in any situation. The road friction is mainly influenced by the road surface and the road condition, e.g., the stopping distance of a vehicle driving 50 km/h on wet cobble with 33.30 m is much larger than on dry asphalt with 21.37 m [25]. In most cases, i.e., moderate outside temperature, the condition is determined by the amount of water on the road section the tires are in contact with.

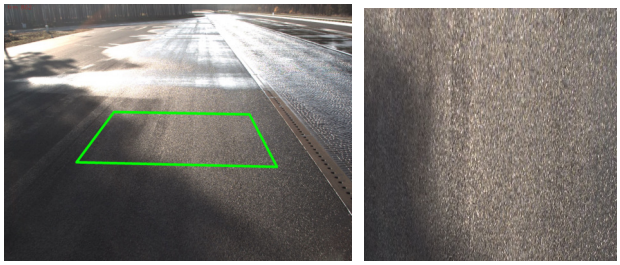
For vehicles, there exists a large variety of approaches



(a) Water film height for each frame f_i



(b) Asphalt, very wet (68 µm), f_{562}



(c) Asphalt, damp (21 µm), f_{862}

Figure 1. The ground truth measurement for the water film height (a) for the selected ROI in the camera images (b), (c). The image patches on the right are derived from the input images (left) using camera calibration. Horizontal gray lines in (a) indicate the proposed split in four levels of wetness {*dry, damp, wet, very wet*}.

Table 1. Comparison of available state-of-the-art datasets for road condition estimation and our proposed *RoadSaW*. The proposed dataset provides controlled conditions on a test track, *Bird's Eye View* (BEV) perspective for *Regions of Interest* (ROI), three surface classes, and accurate MARWIS wetness measurements. Additionally, wetness is split in two and four classes for comparative studies.

Dataset	Image Source	Calibration	Annotation	Label Resolution	# Surfaces	# Wetnesses
Tumen et al. [29]	*Google Street View*	✗	best guess	full image	6	0
Roychowdhury [24]	*YouTube*	✗	best guess	full image	0	4
Busch et al. [8, 22]	compiled	✗	best guess	full image	4	2
RoadSaW (ours)	test track recording	✓	MARWIS	ROI, BEV	3	2/4/∞

for wetness estimation. They use active control systems [32], microphones [6, 18, 23], capacitive [13], optical [1, 2], or vibroacoustical onboard sensors [4]. The *Mobile Advanced Road Weather Information Sensor* (MARWIS) [1] currently provides the most accurate measurements available on the market using optical methods based on the infrared spectrum [27]. However, usage of the MARWIS in commercial vehicles is impossible due to its size and cost. It is an established reference system with no potential for usage in mass production [27].

Current solutions for road condition estimation are only capable of reacting to the actual friction situation. Measurements are taken under or directly in front of the vehicle. In contrast, camera-based approaches, e.g., observing the road from the driver's perspective, are able to make a prediction, i.e., to estimate the condition of a road section before approaching it. This enables the driver assistance system to mitigate the risk of road condition changes before they occur. Camera-based approaches using polarization filters [17] require a stereo configuration which is often not available in current vehicles. A monocular front camera is part of the basic hardware in many vehicles making it a very attractive sensor for road condition estimation.

Since camera images contain high-resolution texture information, it is possible to distinguish between different road surfaces in addition to estimating their condition. Existing approaches have shown that deep convolutional neural networks (DCNN) are capable of making a distinction into two {*dry, wet*} or four {*dry, damp, wet, and very wet*} road wetness classes and provide a classification of the road surface type [8, 9, 21, 22, 25]. Existing datasets collect images from *Google Street View*, *YouTube*, or combine multiple datasets, cf. Tab. 1. However, the annotation is provided as a best guess only. Thus, the comparability of the results is impaired and the generalizability of the approaches is hard to evaluate. A main reason is the difficulty of capturing reasonable video data in combination with ground truth information, such as different levels of wetness, within the driving scenario. The differentiation into multiple degrees of wetness in combination with different road surfaces is of high interest, since these cause significant changes in the driving characteristics. Hence, the objective is to build a dataset with ground truth information from accurate, synchronized measurements using vehicles driving on a test track under controlled surface and wetness conditions.

Contributions We provide a novel dataset, combining ground truth information for water film heights and road surface types. The data is recorded on 10 different days on a test track using cameras mounted on vehicles (truck / car) together with a reference sensor (MARWIS) measuring the water film height. This sensor is currently the best reference system for this task [27] and provides high resolution measurements for the wetness of road regions. The synchronized sensor data (video, MARWIS, and vehicle speed) is captured under controlled conditions using sprinklers on three different road surface types.

Accurate camera calibration enables the construction of bird's eye view road patches with ground truth water film heights and surface types as visualized in Fig. 1. The positions of a road patch and a MARWIS measurement are aligned using the distance traveled by the vehicle, computed by integrating its velocity. We are sure that the proposed dataset is very valuable for future research on road condition and friction estimation. It is the first dataset providing accurate water film height measurements combined with road surface types.

We evaluate the use of deep convolutional neural networks to jointly estimate the combined road surface and road condition. As a baseline, a real-time capable approach using the MobileNetV2 [26] architecture and uncertainty estimation with RBF networks [30] is employed. To investigate the uncertainty estimation, we show evaluations using (a) an out-of-distribution (OoD) dataset with images that do not belong to the input domain as well as (b) close-to-distribution datasets, i.e., real scenes recorded under different conditions, such as lighting (different seasons) and perspective (car instead of truck).

To summarize, our contributions are as follows:

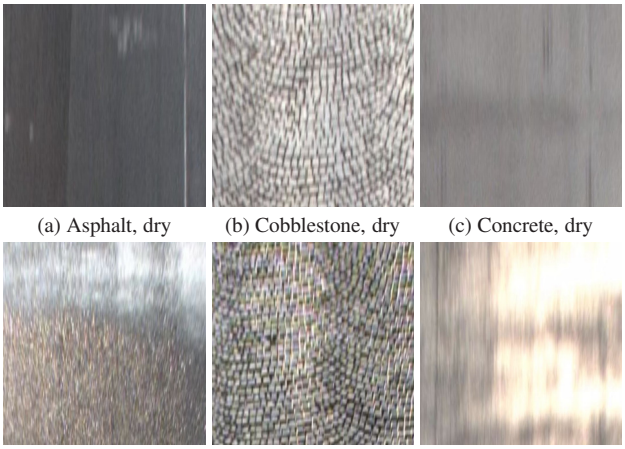
- New dataset for road condition estimation, including accurate water film height measurements combined with road surface types
- Baseline for image-based road condition classification and uncertainty estimation

- Evaluation of close-to-distribution data demonstrating domain changes and generalization performance
- The dataset is available at: https://roadsaw.viscoda.com

2. Related Work

Datasets Due to an increasing usage of machine learning based approaches for automated driving tasks, there is a high demand for datasets. For camera-based road condition estimation, several datasets are used, but only a few of them are suited for achieving reproducible results. These datasets are summarized in Tab. 1. In [29] images are selected from *Google Street View* with different surface types. The approach in [24] uses videos from *YouTube* which are classified into a number of wetness degrees. In [8, 22] images from several existing datasets in combination with the author's own recordings and open-source images are used. In all three cases, the best guess for class annotations for a full image is provided. Regarding the availability of datasets, research of road surface and condition estimation appears to be at a very early stage. For the estimation of combined road surface and wetness, there is no dataset demonstrating its usability in comparable research yet. Thus, algorithms targeting the road surface and condition estimation are currently not reproducible. Our aim is to fill this gap and provide a dataset for comparable research on this important topic.

Uncertainty Estimation While deep neural networks achieve very impressive results in various tasks, the quantification of predictive uncertainty is still a challenging research topic. However, for automated driving the estimation of uncertainty is a crucial topic. Uncertainty is classified into two different categories: aleatoric uncertainty and epistemic uncertainty [3]. Aleatoric uncertainty occurs due to inherent effects inside the data domain such as noise in the input data. Epistemic uncertainty is also known as knowledge or model uncertainty and occurs due to the lack of training data. In the context of automated driving, this type of uncertainty is important since novel situations are unavoidable in real-world scenarios. In this case, the system should react with high uncertainty. Several works have been proposed for uncertainty estimation [3, 5, 14–16, 28]. Bayesian neural networks and *Deep Ensembles* have emerged to be two of the most popular methods [16, 19]. Bayesian neural networks provide a natural way of modelling uncertainty by learning the distribution over weights. However, they require significant modifications to the training procedure and are computationally expensive. A more efficient approach for deterministic neural networks is the usage of *Deep Ensembles* [19]. For many real-time applications, they are still not efficient enough because several forward passes are required. *Deterministic Uncertainty Quantification* (DUQ) as proposed in [30] addresses this issue by learning a *Radial Basis Function* (RBF) network that requires one single forward pass and is just 25% slower at test time than standard classification architectures without uncertainty estimation. DUQ is capable of covering aleatoric and epistemic uncertainty, making it a good choice for the evaluation of our dataset.



(a) Asphalt, dry (b) Cobblestone, dry (c) Concrete, dry

(d) Asphalt, wet (e) Cobblestone, wet (f) Concrete, wet

Figure 2. Examples for road surface patches and their conditions; The data is recorded at various velocities. Often, significant motion blur occurs (b),(c). Depending on the viewing angle, diverse reflections on wet surfaces are present (d),(f).

3. Road Surface and Wetness Dataset

The Road Surface and Wetness (*RoadSaW*) dataset is recorded on a test track using a camera mounted on a truck together with Lufft's MARWIS [1] which measures the water film height. The wetness on the road is induced under controlled conditions using sprinklers on three different surfaces, asphalt, cobblestone (basalt), and concrete. The data is recorded on 10 different days resulting in a large-scale dataset designed for road surface and wetness classification and uncertainty estimation. Camera calibration and data synchronization (MARWIS, video, and velocity) enables the registration of regions of interest to the MARWIS measurement. Various driving maneuvers are performed on all surfaces at different wetness levels. In addition to measurements taken at constant speed, acceleration and deceleration runs are performed, providing speeds of up to $80\,\mathrm{km/h}$. Examples for surfaces in dry and wet condition are shown in Fig. 2. Additionally, *Close-to-Distribution* (CtD) datasets are generated which enable the evaluation of the generalizability of approaches under examination.

3.1. Image Acquisition

The camera (FLIR Blackfly 3.2 MP, 30 fps) is installed behind the windscreen of the truck at 2.66 m height and the car at 1.26 m height. The basis dataset is built upon data recorded on the truck. For the *Close-to-Distribution* dataset as presented in Sect. 3.5 data recorded in the car is used as well. The cameras are carefully calibrated in a preprocessing step using accurately measured 3D landmarks, manually annotated 2D positions of their projections, and minimization of the mapping error as proposed in [10]. The calibration enables the association between camera image and 3D world. It provides the mapping to generate the *Bird's Eye View* (BEV) on the road surface at any visible position (cf. Fig. 1). The temporal synchronization with the data recorded in the vehicle (MARVIS and velocity) is done using timestamps. Within the *RoadSaW* dataset, three different patch sizes ($2.56\,\mathrm{m}^2$, $7.84\,\mathrm{m}^2$, $12.96\,\mathrm{m}^2$) extracted at four different distances to the vehicle (7.5 m, 15 m, 22.5 m, 30 m) are provided. Example patches are shown in Fig. 2. Different patch sizes are considered to evaluate the dependency on the amount of texture required for the estimation. The different distances are useful for the generalization with respect to the reflectivity of the wet surface.

3.2. Water Film Height Measurement

The *Mobile Advanced Road Weather Information Sensor* (MARWIS) [1] detects several road- and weather-related parameters. The MARWIS is mounted at the front of the vehicle, with its measurement unit 1 m above the surface (cf. Fig. 3). Water film heights are measured with a resolution of 1 µm. For the dataset *RoadSaW*[12], the distinction of the wetness into four classes {*dry, damp, wet, very wet*} is provided and combined with three surface types {*asphalt, cobblestone, concrete*}. Additionally, a dataset with a reduced number of classes, *RoadSaW*[6], is provided which employs two wetness degrees {*dry, wet*}, leading to 6 classes. Both sets use the same images. Due to the structure of the surfaces, the maximum possible water film height is only achievable on a concrete surface, leading to different thresholds for each class as shown in Tab. 2.

3.3. Synchronization

In addition to data from the MARWIS and the front camera, the current speed is recorded using the in-vehicle CAN. The temporal synchronization of MARWIS, video, and speed data is established by their timestamps. For the spatial synchronization, i.e., the association of a MARWIS measurement to the respective road surface as visible in an image, the distance traveled by the vehicle is derived from the integration of the speed data for the respective time interval. Thus, the water film height of a region on the road is associated to the respective MARWIS measurement at a later point in time.

Table 2. Overview of the 12 classes included in *RoadSaW*[12]. Derived from the measurements, the assignment of the water film heights has surface dependent thresholds. Some image examples are shown in Fig. 2. *RoadSaW*[6] merges {*dry, damp*} and {*wet, very wet*}.

Class		Water Film Height
Asphalt	dry	0 µm $\leq h <$ 10 µm
Asphalt	damp	10 µm $\leq h <$ 25 µm
Asphalt	wet	25 µm $\leq h <$ 50 µm
Asphalt	very wet	50 µm $\leq h$
Cobble (basalt)	dry	0 µm $\leq h <$ 10 µm
Cobble (basalt)	damp	10 µm $\leq h <$ 25 µm
Cobble (basalt)	wet	25 µm $\leq h <$ 75 µm
Cobble (basalt)	very wet	75 µm $\leq h$
Concrete	dry	0 µm $\leq h <$ 10 µm
Concrete	damp	10 µm $\leq h <$ 60 µm
Concrete	wet	60 µm $\leq h <$ 200 µm
Concrete	very wet	200 µm $\leq h$

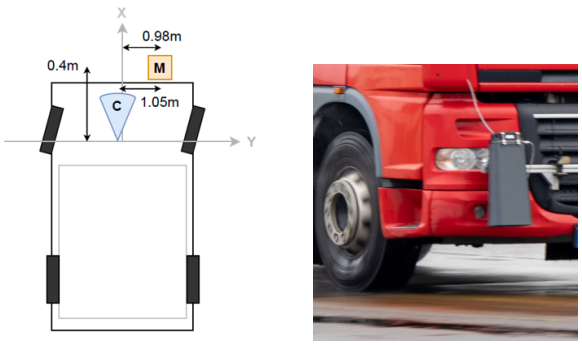


Figure 3. Measurement setup on the truck (left): the orange square M shows the MARWIS and the blue cone C represents the camera. The camera is mounted 2.66 m above the road surface behind the windshield. The mounting position of the MARWIS is outside at the bottom right front of the vehicle (right).

3.4. Dataset Statistics and Balancing

The basis dataset is recorded on five different days ($\approx$ 250 videos) divided nearly equally into *asphalt*, *cobblestone*, and *concrete* with all possible wetness degrees included. From the transformed bird's eye view videos, a *Region of Interest* (ROI) patch is included in the dataset for which a driving distance of 1 m is passed.

Four different distances to the vehicle d2v and three patch sizes (cf. Sect. 3.1) are considered. Overall, there are about 720,000 image patches, each with accurate water film height and velocity measurements. The data is divided into training ($\approx$ 70%), validation ($\approx$ 20%), and test ($\approx$ 10%) sets. Patches recorded at different velocities are evenly distributed among the three sets. Images from the same sequence are assigned to the same set ensuring that similar

CtD^1 (truck) CtD^2 (car) CtD^3 (car)

Figure 4. Example images for the *Close-to-Distribution* datasets. Datasets CtD^2 and CtD^3 are captured with a car (MARWIS visible in images). In the example for CtD^2, wipers occlude main parts of the road.

images are not simultaneously used for both, training and testing.

3.5. Close-to-Distribution Datasets

For validation, *Close-to-Distribution* (CtD) datasets are generated. Like the basis dataset, they are recorded on five different days and consist of synchronized BEV images with ground truth surface and wetness. They provide domain gaps to evaluate the generalizability of approaches for road condition and uncertainty estimation. Compared to *Out-of-Distribution* (OoD) data [19, 30], these datasets are close to the basis dataset with certain domain changes. They fit to the targeted use case and should lead to more interpretable results than OoD data. The data has the following domain changes compared to the basis dataset:

PC *Perspective change*: camera mounted on a different vehicle (car) at lower height ($1.26\,\mathrm{m}$ instead of $2.66\,\mathrm{m}$)

SC *Season change*: data recorded in a different season (May/June instead of November)

TA *Temporary artefacts*: wetness and raindrops on the wind shield, occlusions due to windshield wipers

As shown in Tab. 3, three sets are generated with certain domain changes (PC, SC, TA). For each of the sets $\{CtD^1, CtD^2, CtD^3\}$, eight image sequences of $13\,\mathrm{sec}$ to $33\,\mathrm{sec}$ duration at 30 fps are incorporated. Example images (from driver's perspective) are shown in Fig. 4.

Table 3. Overview of the *Close-to-Distribution* (CtD) datasets: Compared to the basis dataset, perspective change (PC), season change (SC), and temporary artefacts (TA) are provided.

Dataset		PC	SC	TA
CtD^1	Truck 2021-05		✓	
CtD^2	Car 2020-11	✓		✓
CtD^3	Car 2021-06	✓	✓	✓

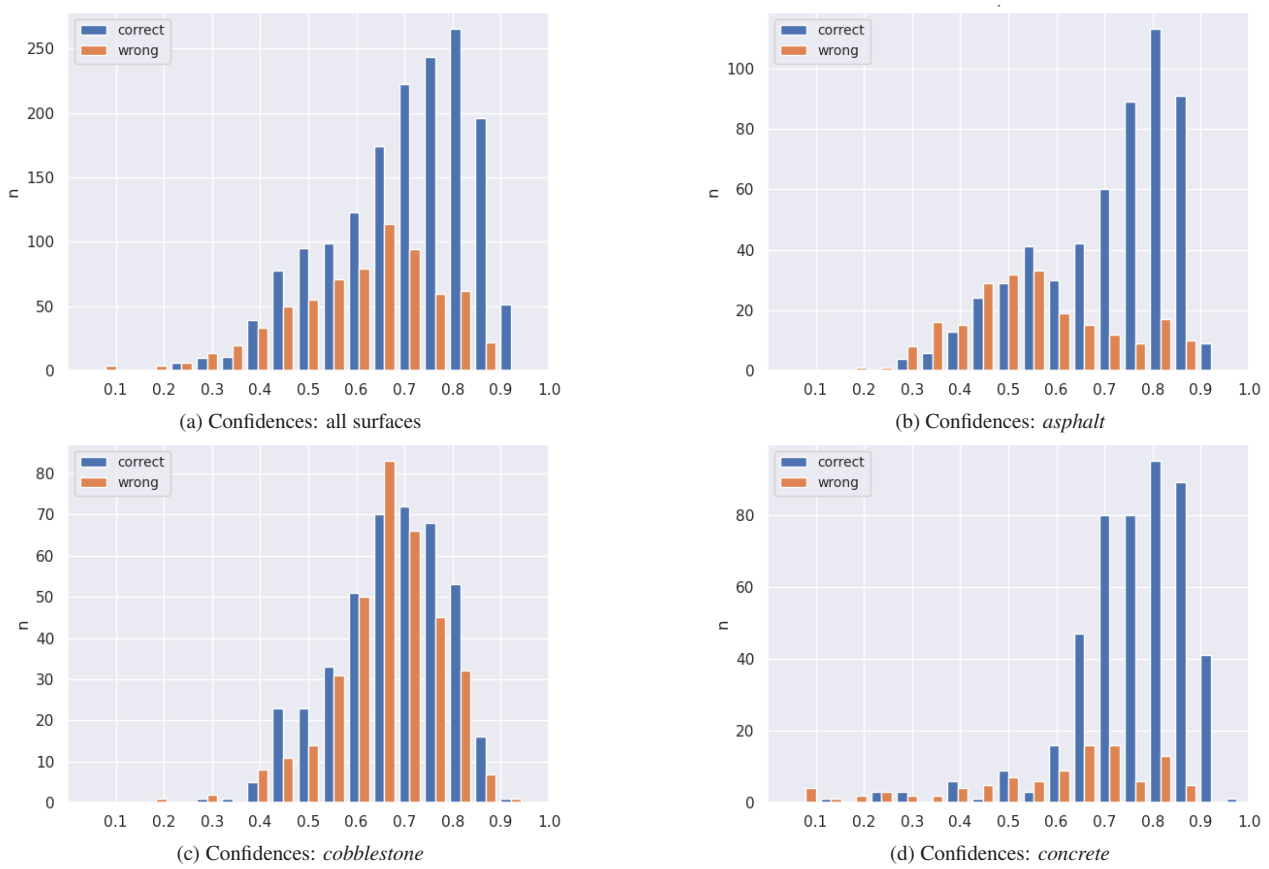
4. Dataset Evaluation

As a baseline, we select state-of-the-art classification algorithms for the evaluation of the *RoadSaW* (*Road Surface and Wetness*) dataset. It is based on the real-time capable architecture MobileNetV2 [26] and the uncertainty estmation using RBF (*Radial Basis Function*) networks [30]. In Sect. 4.1 the evaluation setup is described. The results are presented in Sect. 4.2 (classification) and Sect. 4.3 (RBF). A short discussion concludes this section.

4.1. Evaluation Setup

The targeted use case of automated driving demands minimal inference time such that the onboard system is able to react to the detected situation in a reasonable time. As a good compromise between complexity and inference time, MobileNetV2 [26] is chosen which runs in $\approx 20.5\,\mathrm{ms}$ on a NVIDIA Jetson TX1 [7]. The network is pretrained on *ImageNet* [12]. The architecture is used as feature extractor backbone, followed by a global average pooling layer and a dense layer with softmax as activation function for the classification. All experiments are repeated five times and their mean is reported.

Training For the target architecture MobileNetV2 an image size of 244×244 pixel is used. Due to imbalanced classes, subsampling is performed to obtain an even class distribution. To achieve a better generalization of the trained weights, standard data augmentation is applied during training, i.e., random flipping horizontally, scaling $[90\%, 110\%]$, shifting horizontally and vertically $[-10\%, 10\%]$, and shearing $[-10\%, 10\%]$. The training of the network is done in two steps. First, the weights of the backbone are frozen and the classifier with randomly initialized weights is trained for 10 epochs with learning rate of 10^{-4}. Then, the whole network is fine-tuned for 10 epochs and the learning rate is reduced by a factor of 1000. In both steps the learning rate is reduced by 10% with each epoch. As loss function, the categorical cross-entropy is employed. It is minimized using the RAdam [20] optimizer. For DUQ, two important hyperparameters are the length scale σ and a gradient penalty of λ. The length scale is tuned using the accuracy on the validation set. The gradient penalty is tuned based on the in-distribution uncertainty using the AUROC measure. We use a length scale of $\sigma = 0.1$ and the gradient penalty $\lambda = 0.3$. Here, splitting the training process in two steps is not beneficial. Therefore, the whole network is trained from the beginning with a learning rate of 10^{-4} and is reduced by 10% every epoch. Despite linear decay of the learning rate, the network starts to overfit at a certain point. Early stopping is applied to counteract this.



(a) Confidences: all surfaces

(b) Confidences: *asphalt*

(c) Confidences: *cobblestone*

(d) Confidences: *concrete*

Figure 5. Histograms of confidences as computed with *DUQ* on the dataset *RoadSaW*[12]. The top left (a) shows the histogram for all surfaces, the others (b),(c),(d) show histograms separated for each road surface type. The confidences on *cobblestone* are less accurate than for *asphalt* and *concrete*.

4.2. Results: Classification

The classification performance for different patch distances to the vehicle d2v (cf. Sect. 3) for *RoadSaW*[6] and *RoadSaW*[12] is shown in Tab. 4. The resulting accuracy decreases with increasing distance d2v. Results for different patch sizes (cf. Sect 3.1) are shown in Tab. 5. Larger road sections contain more contextual information, which leads to an improvement of the classification. As expected, the F1-Score for *RoadSaW*[12] is generally smaller compared to *RoadSaW*[6]. The limited accuracy on *RoadSaW*[12] shows that there is potential for a more detailed distinction of wetness classes.

4.3. Results: RBF Networks

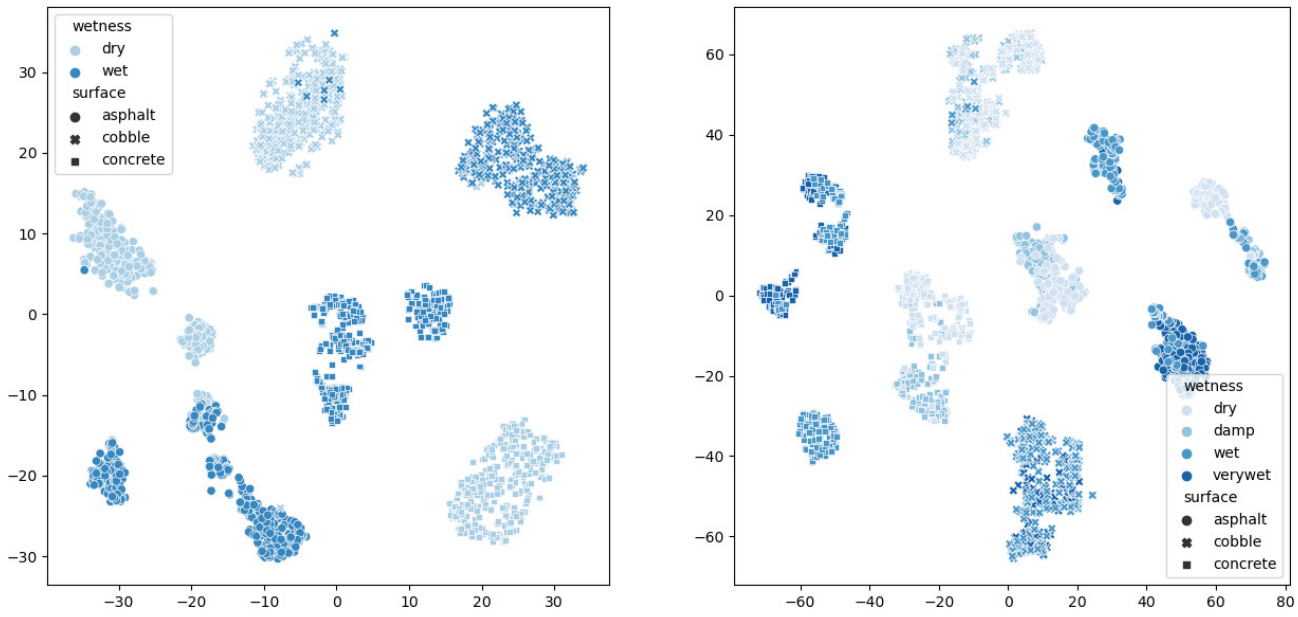
For uncertainty estimation, DUQ (*Deterministic Uncertainty Quantification*) [30] is selected. DUQ is based on *Radial Basis Function* (RBF) networks and proved to provide reasonable uncertainties on available datasets (MNIST, CIFAR, SVHN). Additionally, an *Out-of-Distribution* (OoD) dataset is evaluated [19,30]. As OoD dataset, image patches from Cityscapes [11] are used. This data should receive lower confidence scores than all images from the original dataset. The AUROC (*Area Under the Receiver Opera-*

Table 4. F1-Scores for 6 classes and 12 classes at different **distances to the vehicle d2v** with medium image patch size.

d2v	*RoadSaW*[6]	*RoadSaW*[12]
7.5 m	$91.58\% \pm 0.26$	$64.24\% \pm 0.57$
15.0 m	$90.81\% \pm 0.64$	$61.60\% \pm 0.75$
22.5 m	$84.77\% \pm 0.60$	$57.72\% \pm 0.71$
30.0 m	$80.32\% \pm 0.32$	$58.27\% \pm 0.33$

Table 5. F1-Scores for 6 classes and 12 classes with **different patch sizes** extracted at d2v = 7.5 m distance.

Patch size	*RoadSaW*[6]	*RoadSaW*[12]
Small	$89.56\% \pm 0.47$	$57.43\% \pm 1.12$
Medium	$91.58\% \pm 0.26$	$64.24\% \pm 0.57$
Large	$92.85\% \pm 0.53$	$64.33\% \pm 1.07$



(a) *RoadSaW*^{6} (b) *RoadSaW*^{12}

Figure 6. Visualization of the embeddings of the RBF network DUQ on *RoadSaW* using t-SNE. The three surfaces *asphalt*, *cobblestone*, and *concrete* are represented by the different markers. The four subclasses *dry*, *damp*, *wet*, and *very wet* are visualized with four shades of blue.

tor Characteristic) metric is used to evaluate this property. For the visualization of the resulting cluster configuration, t-SNE (*t-Distributed Stochastic Neighbor Embedding*) [31] is employed. Finally, results on the CtD datasets as introduced in Sect. 3.5 are shown.

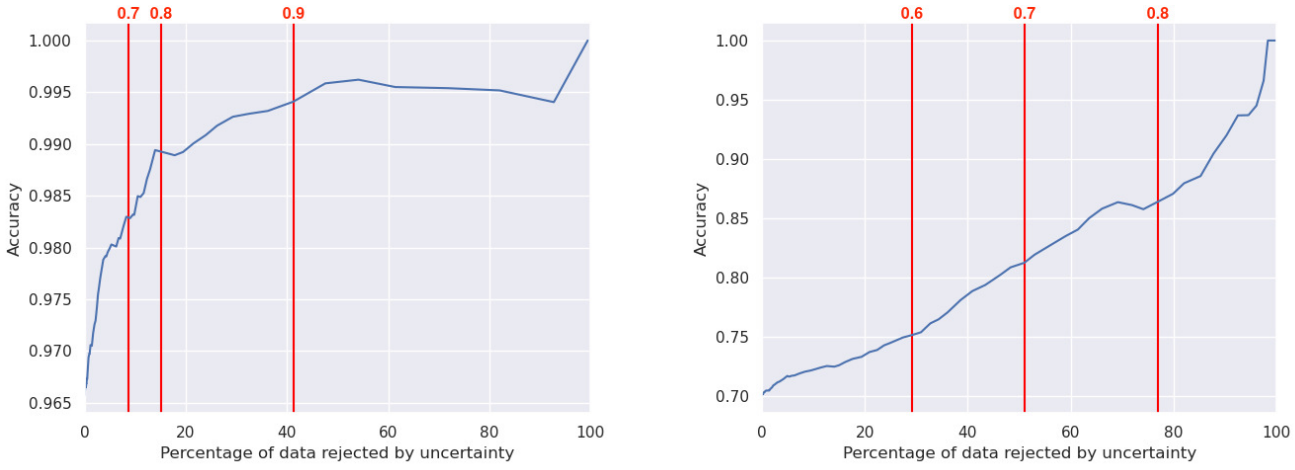
The evaluation (7.5 m d2v, medium patch size) using F1-score leads to $95.77\% \pm 0.8$ on *RoadSaW*^{6} and $70.79\% \pm 1.56$ on *RoadSaW*^{12}. This means an improvement of 6.22 percentage points compared to the results achieved with the standard MobileNetV2 architecture, cf. Sect. 4.2. The results using AUROC for *In-Distribution* and *Out-of-Distribution* data are shown in Tab. 6. For an analysis of the dependencies of the results on the different surfaces, histograms for the confidence estimation on *RoadSaW*^{12} are visualized in Fig. 5. While for *asphalt* and *concrete* (Fig. 5b, 5d), the estimation provides reasonable confidences (mostly correct classifications for confidences close to 1.0), the differentiation between correct and wrong classifications is less accurate for *cobblestone* (Fig. 5c). The analysis using t-SNE visualization in Fig. 6b confirms this observation: the four wetness classes of *asphalt* and *concrete* are well separated while for *cobblestone*, only two distinct clusters are visible. Thus, there are many misclassifications between wetness classes of *cobblestone*. This also degrades the confidence estimation accuracy. For *RoadSaW*^{6} as visualized in Fig. 6a, all six clusters are well-separated leading to high accuracies in classification and uncertainty estimation compared to *RoadSaW*^{12} (cf. Tab. 6). Only a few isolated errors occur.

Table 6. Uncertainty estimation results for the RBF network DUQ. We measure the AUROC score on *In-Distribution* (ID) and *Out-of-Distribution* (OoD) datasets.

Dataset	*RoadSaW*^{6}	*RoadSaW*^{12}
ID	$81.87\% \pm 5.06$	$74.86\% \pm 3.00$
OoD	$98.59\% \pm 0.84$	$96.17\% \pm 3.29$

In many applications, confidences are used to reject uncertain classifications. The required safety level determines the number of rejections. The accuracy as a function of the percentage of rejections is shown in Fig. 7. For *RoadSaW*^{6} (Fig. 7a), the threshold of 0.7 provides an accuracy of 98.3%, rejecting 11% of measurements. For *RoadSaW*^{12} (Fig. 7b), rejecting all classifications with a confidence lower than 0.7, leads to an accuracy of 82%. Then, approximately half of all images are not considered.

The results for the *Close-to-Distribution* (CtD) datasets as described in Sect. 3.5 are shown in Tab. 7. Current models are not able to generalize which leads to insufficient performance on the CtD datasets. The confidences indicate a reasonable trend. This result clearly shows issues on the generalization performance of current approaches. New methods can be developed and analyzed using *RoadSaW*.



(a) *RoadSaW*^{6} (b) *RoadSaW*^{12}

Figure 7. Rejection classification plots of the water film height dataset for 6 classes (a) and for 12 classes (b). The x-axis represents the proportion of data rejected based on the uncertainty score. The red vertical lines indicate the corresponding confidence threshold.

Table 7. **F1-Scores / Mean Confidences** for the *Close-to-Distribution* (CtD) datasets using DUQ classification with 6 classes (*RoadSaW*^{6}) and 12 classes (*RoadSaW*^{12}).

CtD dataset	*RoadSaW*^{6}	*RoadSaW*^{12}
CtD^1	88.28% / 0.76	20.93% / 0.44
CtD^2	28.88% / 0.58	15.94% / 0.47
CtD^3	51.18% / 0.54	0.78% / 0.37

4.4. Discussion

The experiments demonstrate the usability of the proposed *RoadSaW* dataset in the context of classification using 6 classes (*RoadSaW*^{6}) or 12 classes (*RoadSaW*^{12}). The class selection is inspired by state-of-the-art road condition estimation approaches [8, 24, 25, 29]. The evaluations and visualizations are aligned with approaches in the field of classification and uncertainty estimation such as [19,30,31]. The results show that the selected approaches solve the problem for *In-Distribution* (ID) data with reasonable accuracy. However, for the class *cobblestone* the classification appears more difficult when four levels of wetness are used. We also tested standard regression approaches to predict the exact water film heights, but were not able to improve the classification accuracy after their mapping to the respective classes. There is substantial potential for accuracy improvements on *RoadSaW*^{12}.

The experiments on CtD data show that the generalization to data with certain domain gaps is an open problem. New augmentation or self-supervised learning approaches [33] are promising for the improvement of the feature representation. *RoadSaW* provides the data for future improvements and evaluations.

5. Conclusion

Although the application of road friction estimation for automated driving is an important topic, research on camera-based road surface condition estimation is at an early stage. Currently, there is no dataset available for the comparative study and evaluation of road condition estimation approaches. Thus, a new dataset is proposed targeting the combined *Road Surface and Wetness* estimation, called *RoadSaW*. It includes 720,000 bird's eye view patches recorded on a test track together with high resolution video and synchronized, accurate water film height measurements. As surface types, *asphalt*, *concrete*, and *cobblestone* are included. The dataset is evaluated using state-of-the-art, real-time machine learning based approaches for classification and uncertainty estimation. While reasonable uncertainties are achieved on *In-Distribution* (ID) and *Out-of-Distribution* (OoD) data, the performance on the provided *Close-to-Distribution* (CtD) datasets is insufficient. For the application of road condition estimation, more robustness is required. We would like to create the possibility to develop and benchmark new methods which is important for the research community working on automated driving.

Acknowledgements This work has been performed in the framework of the *InFusion* project supported by the *Bundesministerium für Verkehr und digitale Infrastruktur*. The authors would like to acknowledge the contributions of their colleagues from *InFusion*.

References

[1] Lufft: MARWIS (Mobile Advanced Road Weather Information Sensor). https://lufft-marwis.de/en/specifications/. [Online; accessed 09-Mar-2022]. 2, 3, 4

[2] Optical sensors: Roadeye. http://www.opticalsensors.se/roadeye.html. [Online; accessed 09-Mar-2022]. 2

[3] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021. 3

[4] I. Abdić, L. Fridman, D. E. Brown, W. Angell, B. Reimer, E. Marchi, and B. Schuller. Detecting road surface wetness from audio: A deep learning approach. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3458–3463. IEEE, 2016. 2

[5] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020. 3

[6] S. Bahrami, S. Doraisamy, A. Azman, N. Amelina Nasharuddin, and S. Yue. Acoustic feature analysis for wet and dry road surface classification using two-stream cnn. In *2020 4th International Conference on Computer Science and Artificial Intelligence*, pages 194–200, 2020. 2

[7] S. Bianco, R. Cadene, L. Celona, and P. Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018. 5

[8] A. Busch, D. Fink, M.-H. Laves, Z. Ziaukas, M. Wielitzka, and T. Ortmaier. Classification of road surface and weather-related condition using deep convolutional neural networks. In *The IAVSD International Symposium on Dynamics of Vehicles on Roads and Tracks*, pages 1042–1051. Springer, 2019. 2, 3, 8

[9] S. R. Chowdhury, M. Zhao, M. Jonasson, and N. Ohlsson. Methods and systems for generating and using a road friction estimate based on camera image signal processing, July 7 2020. US Patent 10,706,294. 2

[10] K. Cordes and H. Broszio. Vehicle lane merge visual benchmark. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 715–722, 2021. 4

[11] 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 (CVPR)*, 2016. 6

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[13] J. Döring, L. Tharmakularajah, J. Happel, and K.-L. Krieger. A novel approach for road surface wetness detection with planar capacitive sensors. *Journal of Sensors and Sensor Systems*, 8(1):57–66, 2019. 2

[14] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021. 3

[15] S. Farquhar, M. A. Osborne, and Y. Gal. Radial bayesian neural networks: beyond discrete support in large-scale bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1352–1362. PMLR, 2020. 3

[16] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 3

[17] M. Jokela, M. Kutila, and L. Le. Road condition monitoring system based on a stereo camera. In *2009 IEEE 5th International conference on intelligent computer communication and processing*, pages 423–428. IEEE, 2009. 2

[18] M. Kalliris, S. Kanarachos, R. Kotsakis, O. Haas, and M. Blundell. Machine learning algorithms for wet road surface detection using acoustic measurements. In *2019 IEEE International Conference on Mechatronics (ICM)*, volume 1, pages 265–270. IEEE, 2019. 2

[19] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 3, 5, 6, 8

[20] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 5

[21] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016. 2

[22] M. Nolte, N. Kister, and M. Maurer. Assessment of deep convolutional neural networks for road surface classification. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 381–386. IEEE, 2018. 2, 3

[23] G. Pepe, L. Gabrielli, L. Ambrosini, S. Squartini, and L. Cattani. Detecting road surface wetness using microphones and convolutional neural networks. In *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019. 2

[24] S. Roychowdhury, M. Zhao, A. Wallin, N. Ohlsson, and M. Jonasson. Machine learning models for road surface and friction estimation using front-camera images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 2, 3, 8

[25] E. Šabanovič, V. Žuraulis, O. Prentkovskis, and V. Skrickij. Identification of road-surface type using deep neural networks for friction coefficient estimation. *Sensors*, 20(3):612, 2020. 1, 2, 8

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2, 5

[27] B. Schmiedel, F. Gauterin, and H.-J. Unrau. Road wetness quantification via tyre spray. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of automobile engineering*, 233(1):28–37, 2019. 2

[28] M. Teye, H. Azizpour, and K. Smith. Bayesian uncertainty estimation for batch normalized deep networks. In *Interna-*

tional Conference on Machine Learning, pages 4907–4916. PMLR, 2018. 3

[29] V. Tumen, O. Yildirim, and B. Ergen. Recognition of road type and quality for advanced driver assistance systems with deep learning. *Elektronika ir Elektrotechnika*, 24(6):67–74, 2018. 2, 3, 8

[30] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020. 2, 3, 5, 6, 8

[31] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008. 7, 8

[32] M. Wielitzka, M. Dagen, and T. Ortmaier. State and maximum friction coefficient estimation in vehicle dynamics using UKF. In *2017 American Control Conference (ACC)*, pages 4322–4327. IEEE, 2017. 2

[33] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 8