Road Surface Segmentation - Pixel-Perfect Distress and Object Detection for Road Assessment

Ronny Stricker[a], Dustin Aganian[a], Maximilian Sesselmann[b], Daniel Seichter[a], Marius Engelhardt[a],
Roland Spielhofer[c], Matthias Hahn[c], Astrid Hautz[d], Klaus Debes[a], and Horst-Michael Gross[a]

[a] Technische Universität Ilmenau
Neuroinformatics and Cognitive Robotics Lab
98684 Ilmenau, Germany
ronny.stricker@tu-ilmenau.de

[b] LEHMANN + PARTNER GmbH
99086 Erfurt, Germany

[c] AIT Austrian Institute of Technology GmbH
Center for Low-Emission Transport, 1210 Vienna, Austria

[d] VIA IMC GmbH
Franz-Ehrlich-Str. 5, 12489 Berlin

***Abstract*—Visual road assessment, which is carried out by many countries, involves the evaluation of millions of surface images. This exhaustive task is usually done manually and therefore is costly in terms of time and prone to failure. Different methods for automatic distress detection have been presented in the literature recently. However, most of the approaches are focused on crack detection only. This paper focuses on detecting multiple distress types and object classes on asphalt roads, aiming to fully automate distress detection on road surfaces in Austria, Switzerland, and Germany using image segmentation with neural networks. The paper introduces a distress and object catalog developed by experts of the involved countries that guarantees convertibility into federal distress catalogs. We evaluate the performance gain of different neural network architectures and advanced training techniques by conducting extensive experiments.**

***Index Terms*—distress detection, segmentation, asphalt pavement**

I. Introduction

The public road infrastructure is constantly aging and needs frequent inspections to guarantee its permanent availability. Following the federal regulations of Austria, Switzerland, and Germany, federal roads' surface characteristics have to be evaluated regularly, i.a. regarding substance conditions. The substance condition describes the visible part of the surface characteristics. It is evaluated by visual inspection of surface images recorded with the help of mobile mapping systems (Fig. 1). Even though automatic image processing has been applied to various application domains, evaluating these images is done manually and requires excessive manual labor. This process is very time-consuming and exhausting, which leads to inconsistent and faulty distress detection. The manual evaluation also delays the assessment process considerably and can take up to several months. Consequently, the results are already outdated once the assessment has finished.

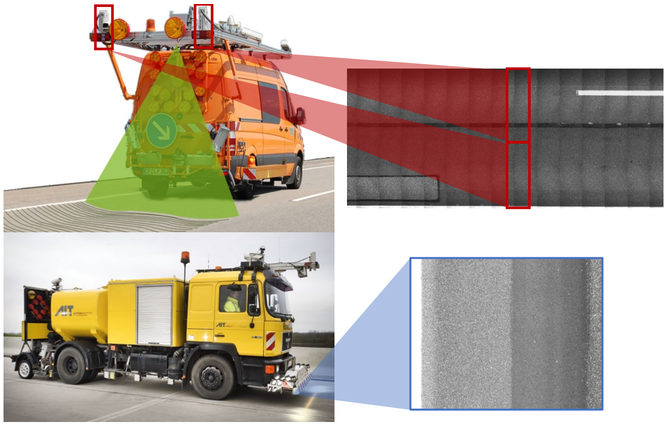


Fig. 1: The mobile mapping systems S.T.I.E.R. (top left) and RoadSTAR (bottom left) are used for capturing road surface images in Austria, Switzerland and Germany. While RoadSTAR is using a line scan camera, S.T.I.E.R. is equipped with two shutter-based cameras. Therefore, images have to be stitched to generate larger surface images.

In the research project ASFaLT[1], we aim to tackle the problems emerging from manual road image assessment with the help of deep-learning-based machine vision. In our previous works [1]–[3], we have already shown that deep neural networks are perfectly capable of road distress detection.

In contrast to our previous work, which was primarily focused on distress detection as a binary decision problem, this work aims to detect all distress classes of relevance for visual road assessment in Austria, Switzerland, and Germany. Therefore, we present a distress catalog for road damages and objects that can be converted into the three countries' federal regulations. Afterward, we show the challenges of the application domain and analyze the benefit of different augmentation and training techniques to improve detection

This work has received funding from Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft or FFG) as part of the ASFaLT project under grant agreement no 869514.

[1] ASFaLT: **A**utomatisierte **S**chadstellenerkennung für unterschiedliche **Fa**hrbahnbeläge mittels Deep **L**earning **T**echniken (Automated distress detection for different road surfaces using deep learning)

performance. We conclude by comparing the results achieved with the system presented in this paper to images manually labeled by different field experts.

In addition to the German Asphalt Pavement Distress (GAPs) dataset [1] and the extended and refined version presented in [2] that provides high-quality standardized images and attracted much attention by several research groups (e.g. [4]–[6]), we provide the publicly available *GAPs 10m* dataset[2] with this paper. The dataset consists of 20 high-resolution images (5030 x 11505 pixels corresponding to 10 meters of the road surface) that cover 200 meters of asphalt roads with different asphalt surface types and a wide variety of distress classes. All images are captured using recording vehicles following German and Austrian federal regulations (see Fig. 1) and feature a high spatial resolution close to one pixel per millimeter. The images are labeled by experts in the field and can be used freely for evaluation purposes.

II. Related Work

A wide variety of different approaches for automatic distress detection has been presented since the first attempt on automatic visual distress detection [7]. Classic image processing techniques have been developed by combining preprocessing algorithms for illumination independence with different thresholding techniques to extract local minima as crack candidates [8]–[12]. Due to the tremendous success of deep-learning-based image processing in almost all image processing domains, manually designed features have fallen out of favor. The algorithms developed for deep learning-based evaluation of the pavement surface can be divided into the following major groups (1-3). An overview of the different publicly available datasets is given at the end of this section (4). Approaches based on Convolutional Neural Networks (CNNs) mainly differ regarding network architecture, predicted distress classes, and whether downward or frontal-facing input images are processed.

1) Crack Detection: The first attempts for CNN based crack detection in [13] and [14] are using LeNet-5 [15] based or VGG-based [16] CNNs for patched based crack detection. Both approaches need to be converted to fully convolutional networks to obtain an image segmentation result. Thus, the use of network structures for image segmentation is becoming more common recently. U-Net [17]-based architectures, which are also quite common in the biomedical image segmentation domain, have been successfully used for crack detection by several researchers [5], [18], [19]. Also, adapted versions of SegNet have been applied in that domain [20].

2) Distress Detection in Orthoframes: The vast majority of research papers are focused on crack detection only. However, [21] applied a U-Net-like network architecture with different context resolution levels to integrate more context. The authors consider different types of distress but classify into distress and normal area only. An approach distinguishing between cracks, sealed cracks, and potholes can be found in [22]. Therefore, the authors used a combination of U-Net and the YOLO approach.

3) Distress Detection in frontal-facing Images: Distress detection in frontal-facing images is often carried out in two stages. The road area is detected by traditional image segmentation techniques like graph-based hierarchical clustering [23] or using CNNs like SegNet [24] in the first stage. The network architectures used for detecting road distress on the extracted road area in the second stage are based on state-of-the-art image processing networks. [25] compares InceptionV2 and MobileNet for the detection of eight different distress classes while [24] applies Squeeze-Net for distress detection. [26] presents a Feature Pyramid and Hierarchical Boosting Network.

This paper is focused on processing orthoframes that are also required for standardized road assessment in Austria, Switzerland, and Germany. Furthermore, approaches working on frontal facing images often concentrate on severe damages that are uncommon on federal roads of the involved countries.

4) Datasets: Although many different methods have been presented so far, there is still a lack of publicly available datasets that are sufficiently large and are recorded in a standardized way. The datasets published so far do often consist of less than 500 images, e.g., [8], [10], [26], [27], and do not offer the necessary diversity to train a universal pavement distress detector. Although some datasets have been released recently that do offer a decent size, e.g. [28], with 700k Google Street View images or [29] with 13k frontal facing images, these datasets are using frontal-facing images. Furthermore, they do not provide the level of resolution required for standardized road assessment and mostly show images with severe distress only.

III. Novel Standardized Distress Dataset

The datasets publicly available do not cover all the different classes needed for distress detection on the level required by federal regulations in Austria, Switzerland, and Germany. Therefore, the data used for road assessment in the involved countries have been brought together within the research project's scope ASFaLT.

A. Standardized Data Acquisition

The data acquisition within the ASFaLT project is based on the specification of the Road Monitoring and Assessment of the countries Germany [30], and Austria [31], [32]. The image data of the dataset have been captured by the mobile mapping system S.T.I.E.R (Fig. 1), one of the systems certified annually by the German Federal Highway Research Institute (BASt) and by the mobile mapping system RoadSTAR [33], mainly deployed in Austria. The vehicle S.T.I.E.R. is equipped with several high-resolution cameras, i.a. two slightly overlapping bird-eye-view photogrammetrically calibrated monochrome cameras capturing the pavement's surface in detail. The surface camera system is synchronized with

[2] The *GAPs 10m dataset* is available at: https://www.tu-ilmenau.de/neurob/data-sets-code/gaps

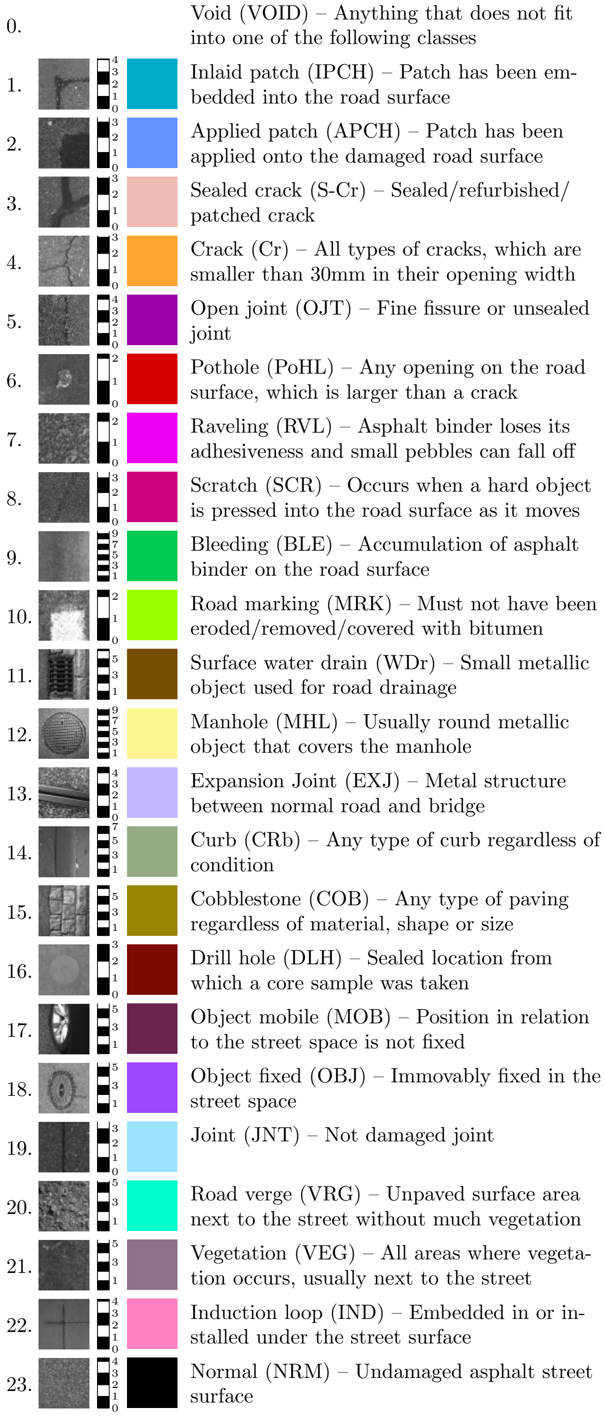
a high-performance lighting unit to reduce ambient light's influence. The mobile mapping system RoadSTAR, in turn, uses a line scan camera in combination with permanent lighting devices. Both mapping systems allow continuous capturing of road surface images even at high velocities (about 80 km/h). For more details regarding the data acquisition process and the measurement vehicle, we refer to [1] and [33]. The mapping devices' images are transformed to have a fixed metric correspondence to comply with federal regulations. Therefore, standardized images in Austria correspond to a metric resolution of 4 meters in width and approx. 3 meters in height. In Germany, however, the images exhibit a height corresponding to 10 meters of road. Since the cameras of the mapping system S.T.I.E.R. do only capture a smaller road fraction, the images recorded have to be stitched to comply with federal regulations (Fig. 1). Both systems differ slightly in respect to the physical area captured by a single image pixel. While RoadSTAR images are captured with roughly 900 pixels per meter, the images of S.T.I.E.R. exhibit a resolution of approximately 1200 pixels per meter. Both systems can carry out the data acquisition in Switzerland. Therefore, no other mapping system has to be considered for that country. Since image acquisition is carried out by different companies with different mobile mapping systems in Germany, the dataset also includes ten images from another mapping system. These images should be used for testing to see how detection generalizes with different mapping systems.

B. *Unified Object Catalog and Labeling*

Although the road surface assessment is carried out similarly in all three countries, the distress classes and the associated metrics are different. Therefore, unified distress classes have to be defined in a first step. The resulting catalog covers a wide variety of distress- and object classes and can be mapped to the federal road assessment regulations of the involved countries. Various additional object classes have been defined that are not used in the current assessment process but might be of future interest (Fig. 2). All dataset images have been labeled with pixel-level accuracy by several trained annotators to guarantee high-quality labels.

C. *Dataset Size*

Following our previous work presented in [3], we opted for an iterative dataset creation process to minimize labeling effort. Starting with a small initial dataset, uncertainty estimation on unlabeled images has been used to select images worth labeling. The final dataset consists of 193 surface images following Austrian federal regulations (4m x 3m images) and 201 images following German federal regulations (10m x 4.5m images). The images are recorded at 99 road sections (groups) that are at least 10 km apart to offer a wide variety of surface characteristics. The dataset has been split into training, validation, and test such that a single group only occurs in one of the subsets. Since the complete dataset cannot be made publicly available for legal reasons, we selected a subset *VAL(p)* of the validation dataset published together with this paper for reproduction and comparison. This smaller subset consists of 20 images following German federal regulations and covers various distress- and object types. The class distribution of the different subsets can be seen in Tbl. I.



0. Void (VOID) – Anything that does not fit into one of the following classes
1. Inlaid patch (IPCH) – Patch has been embedded into the road surface
2. Applied patch (APCH) – Patch has been applied onto the damaged road surface
3. Sealed crack (S-Cr) – Sealed/refurbished/patched crack
4. Crack (Cr) – All types of cracks, which are smaller than 30mm in their opening width
5. Open joint (OJT) – Fine fissure or unsealed joint
6. Pothole (PoHL) – Any opening on the road surface, which is larger than a crack
7. Raveling (RVL) – Asphalt binder loses its adhesiveness and small pebbles can fall off
8. Scratch (SCR) – Occurs when a hard object is pressed into the road surface as it moves
9. Bleeding (BLE) – Accumulation of asphalt binder on the road surface
10. Road marking (MRK) – Must not have been eroded/removed/covered with bitumen
11. Surface water drain (WDr) – Small metallic object used for road drainage
12. Manhole (MHL) – Usually round metallic object that covers the manhole
13. Expansion Joint (EXJ) – Metal structure between normal road and bridge
14. Curb (CRb) – Any type of curb regardless of condition
15. Cobblestone (COB) – Any type of paving regardless of material, shape or size
16. Drill hole (DLH) – Sealed location from which a core sample was taken
17. Object mobile (MOB) – Position in relation to the street space is not fixed
18. Object fixed (OBJ) – Immovably fixed in the street space
19. Joint (JNT) – Not damaged joint
20. Road verge (VRG) – Unpaved surface area next to the street without much vegetation
21. Vegetation (VEG) – All areas where vegetation occurs, usually next to the street
22. Induction loop (IND) – Embedded in or installed under the street surface
23. Normal (NRM) – Undamaged asphalt street surface

Fig. 2: A listing of all classes in the dataset. Each entry shows a sample image, a scale in decimeters, followed by the color of the visualization, the class name, and a short description.

TABLE I: Class distribution of the dataset. The dataset has been split according to different location groups to guarantee that the same road sections are only present in a single dataset. An explanation of the class names can be taken from Fig. 2.

Set	#Grps	IPCH	APCH	S-Cr	Cr	OJT	PoHL	RVL	SCR	BLE	MRK	WDr	MHL	EXJ	CRb	COB	DLH	MOB	OBJ	JNT	VRG	VEG	IND
TRAIN	53	181	320	291	3855	110	281	172	22	135	824	35	27	11	76	225	48	28	26	167	201	332	14
VAL	19	39	8	45	315	3	22	2	7	2	86	3	0	1	9	13	3	3	2	17	26	18	0
VAL(p)	11	10	3	24	191	2	11	0	7	0	47	2	0	1	5	2	3	1	2	6	16	13	0
TEST	27	61	30	23	366	20	47	49	2	8	143	6	4	2	13	29	1	4	3	14	27	51	2

D. *Challenges*

Although the road is captured under controlled conditions, resulting in high-quality surface images, the dataset comprises some challenges. Despite the fact that the mobile mapping systems use artificial light sources, harsh sunlight can introduce image artifacts (Fig. 3) that look similar to applied patches. Actual patches, however, can be hard to detect since their appearance can be very similar to the surrounding asphalt pavement (Fig. 3). Patches induce another challenge since the inside of a patch appears almost identical to intact asphalt regions. Inlaid patches can cover several meters in width and height, making it impossible to distinguish between the patch and normal asphalt regions from a single image. Therefore, patch areas are encoded using their transition area between regular asphalt and patch only.

Another challenge originates from the image stitching process required if the mapping system does not use a line camera. Since different types of mapping systems are used in Germany, the artifacts induced by stitching can differ and hamper classification.

IV. Network Architecture

We analyzed different network architectures on our dataset to determine which network design works best with the suggested class catalog. We tested U-Net [17], which is often used in crack detection, and a U-Net variant that uses Xception [34] blocks. Furthermore, we are applying a network structure that shares similarities with the classic PSPNet [35]. The network has a typical encoder-decoder structure. As the default encoder, we use Residual Networks (ResNet) [36] with dilated convolutions as presented in [37] and with 18 and 50 layers. The context module of the segmentation network is represented by the Pyramid Pooling Module (PPM) [35], or the Atrous Spatial Pyramid Pooling (ASPP) [38]. The final segmentation output is produced by a PSPNet decoder [35]. An overview of our final model is presented in Figure 4.

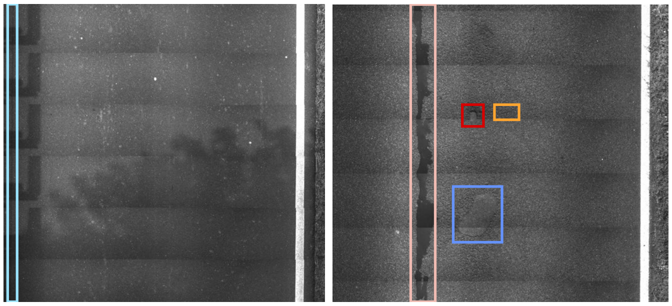


Fig. 3: Bounding boxes for various distress objects (True labels are at pixel level).
left – shadows from the object carrier (next to the joint (cyan)) are similar to an applied patch.
right – Patch in the center of the image is difficult to detect.

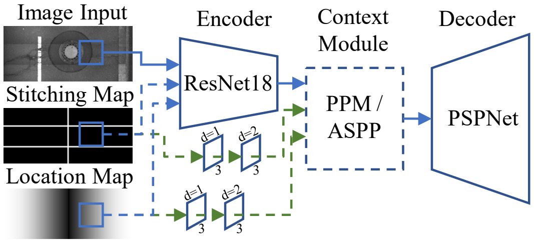


Fig. 4: Overview of our inputs and our model. The stitching map and the location map can optionally be applied as input to the encoder (blue) or to the context module (green).

1) Location map: Some classes in our dataset are hard to detect using local information only and might benefit from additional context. In [39] it has been shown that explicit coding of location information can improve location context. Therefore, we generated a location map, which equals zero in the center of the images and increases linearly to the left and right border of the road image (Fig. 4). The map is integrated into the model in two different ways. First, it can be presented as an additional input channel. Second, it can be integrated directly into the context module. To do that, we have scaled the map to the corresponding context module feature map size and prepended two Convolution-ReLU Blocks (filter size 3, dilation rates 1,2).

2) Stitching map: A second problem arises from the stitching edges in the images produced by the mobile mapping system S.T.I.E.R. since they can easily mix up with patches or joints. Therefore, we added a *stitching map* as a further context that is derived from classic image processing. To detect the vertical stitching edges we used the algorithm we presented in [3]. We apply a vertically oriented Sobel filter to the entire image and sum up the results within every row to obtain a one-dimensional edge candidate vector. This vector is transformed into the frequency domain using a discrete fast Fourier transformation. After maximum filtering, the peaks are back-projected and used to generate a stitching edge map (Fig. 4). Since the horizontal stitching edges are always located at the image center, they do not have to be detected separately.

V. Experiments

This section analyzes the influence of different data augmentation methods, network architectures, and training algorithms on the classification performance to outline techniques worth examining by other researchers working in the road distress detection domain.

A. Occurrence Score

The intersection over union (IoU) can measure the overlap between ground truth and predicted object areas and, therefore, is commonly used to evaluate the segmentation quality among various research areas. However, when it comes to failure detection, it is even more important to measure if the failure can be detected at all. Therefore, we introduced another measure inspired by the German road assessment standard, where distress is assigned to grid cells with a cell size of approximately 1x1 meters. Following that procedure, we divide the image into grid cells with a size of 1x1 meters and observe if a class is detected in that area only. We compare the detection to the ground truth to compute true positives (TP), false positives (FP), and false negatives (FN) for every cell of an image. Afterward, we calculate the critical success index as $\frac{TP}{TP+TN+FP}$ for every single image and take the average over all images as *Occurrence Score* (Occ).

Small false detections can heavily affect the occurrence score. Since these small detections can be filtered in post-processing, we count class detections only if the pixel count of an object exceeds half of the area of the smallest entity of that type in the training dataset (e.g., potholes are filtered if the area is smaller than 38 pixels).

B. Test setup

All models were trained for 300 epochs. Images were scaled to half resolution to evaluate various methods and techniques in a shorter period. For training, random non-overlapping patches with a size of 480x480 pixels were drawn from the dataset in each epoch. The patches were grouped into batches of size 8, and training was performed with the SGD optimizer using multiple learning rates (0.005, 0.01, 0.02, 0.04). A logarithmically decreasing learning rate was used[3]. Networks were randomly initialized for training since ImageNet pre initialization is not beneficial in the surface image domain, as shown in [2]. Class weights were used for training and were determined according to [40]. All results are presented with occurrence scores and mean intersection over union (mIoU) as indicators for segmentation quality. If training was unstable, the training was repeated several times, and the standard deviation is given in parentheses after the best result.

C. Baseline

The baseline experiments have been conducted with the ResNet18 encoder and the PPM context module. Without data augmentation, the network can reach an occurrence score of **Occ 0.61 (0.004)** and a **mIoU of 0.574 (0.003)**.

D. Data Augmentation

One of the most straightforward data augmentation techniques is to flip the image patches horizontally and vertically, which improves the performance to **Occ 0.632 (0.007)** and a **mIoU of 0.586 (0.004)**. Since this technique improves performance for all classes, it is applied to all further experiments. Further different standard data augmentation techniques have been tested one by one to analyze their effect on the network training. We have modified brightness, contrast, noise, patch rotation, and patch scaling by random factors drawn randomly from a uniform distribution with varying upper bound (Fig. 5). Since the input images always contain a multitude of different classes, augmentation has been applied to the whole image regardless of the contained classes.

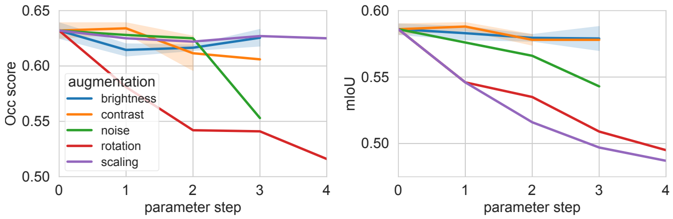


Fig. 5: Comparison of different augmentation techniques. The parameter steps are as follows: brightness - [0, 0.1, 0.2, 0.3], contrast - [0, 0.1, 0.2, 0.3], noise - [0.01, 0.02, 0.03], rotation - [0°, 22°, 45°, 67°, 90°], scaling - [1.0, 0.95, 0.9, 0.85, 0.8]

The augmentation techniques under study, however, did not show any significant performance improvement. Most of them have proven to hamper good segmentation results. This result is quite specific to the distress detection domain with well-controlled image acquisition conditions. If, for example, rotation is introduced, the network is no longer able to learn that some structures like joints do have a certain alignment on the road. Brightness and noise also do not improve classification results. Slight improvements could be achieved with contrast augmentation of 0.1, which, in combination with image flipping, builds the new baseline for the experiments in (E)-(H): **Occ - 0.634 (0.005)**, **mIoU - 0.588 (0.003)**

E. Network Architecture

Besides our baseline, we examined other architectures as described in Section IV and summarized the results in Tbl. II. We found that the U-Nets, which are often used for crack detection, perform well in the crack classes but far worse in all other classes, giving a poor overall result. The far more computationally expensive networks with ResNet50 as the encoder or ASPP as context module have not yielded any noticeable gain over our baseline architecture and are not further examined.

F. Additional context

In Sec. IV we have proposed two context map types that might help the network to involve further context information. Tbl. II reveals that location information does not have a positive effect and even deteriorates results if used directly as network input. One reason might be that classes that may benefit from that location information are already well detected (e.g., water drain).

[3] Starting with the maximum learning rate, the learning rate reaches a minimum of 0.0001 in the last epoch. Additionally, a warmup of 5 epochs at the beginning of the training was applied that scaled linearly from the minimum learning rate to the maximum learning rate.

TABLE II: Scores of experiments (E)-(G)

Experiment	*Occ*	*mIoU*	*Experiment*	*Occ*	*mIoU*
(E) U-Net	.450	.338	(F) Loc. + Ctx.	.637	.577
(E) U-Net (Xception)	.499	.531	(F) Loc. + Inp.	.614	**.598**
(E) ResNet50	**.633**	**.585**	(F) Stitch + Ctx.	**.662**	.585
(E) ASPP	.630	.584	(F) Stitch + Inp.	.640	.589
(G) Cat. cross-entropy	.634	0.588			
(G) Focal loss	.628	.590			
(G) Tversky + Focal	**.646**	**.595**			

In contrast, the stitching context map generated a significant occ score improvement when applied to the context module directly. Slim classes like Crack, Joint, and Scratch do benefit the most. Furthermore, this modification positively influenced the images of the mapping system that were not contained in the training data. False detections on stitching edges were reduced significantly for these images.

G. *Loss functions*

Since the input always comprises a larger road section, we can only work with highly unbalanced data. Therefore, it is reasonable to experiment with other appropriate loss functions besides categorical cross-entropy. These loss functions were Focal Loss [41] and Tversky Loss [42].

During a street segmentation model training, simple classes such as normal, biomass, road marking, etc., occur frequently and are easy to learn. We used the Focal Loss function to prevent the training loss from dropping too quickly after the model has mastered the simple examples. Focal Loss introduces a γ-parameter for the categorical cross-entropy, which can be used to control the well-predicted classes' penalty.

Another problem is that regular asphalt often appears to have tiny cracks, so training can lead to models that tend to detect false positives in the crack class. To solve this problem, Tversky Loss is applied, which is calculated using the Tversky Index and is a generalization of the Dice Coefficient and the Jaccard Index. Using Tversky Loss, false positives or false negatives can be penalized to varying degrees. Tbl. II presents that both loss functions have a positive effect on the occurrence score but that the gains for the given scenario are minimal.

H. *Label smoothing*

As we will show in Sec. VI, the labeling of the roads by experts is a challenging undertaking, where experts often disagree. Therefore, wrong labels are a significant problem constraining the training. Label smoothing [43], that regularizes the classifier layer has been applied to counter this problem. Our experiment can confirm this assumption and yields significantly better results compared to our baseline: **Occ - 0.696**, **mIoU - 0.613** with best $\epsilon = 0.2$.

Mixup [44] has been reported to work even better than label smoothing by blending two inputs and the associated labels with a blending factor drawn from a symmetric Beta distribution $Beta(\alpha, \alpha)$. Although mixup significantly improves the detection (Occ - 0.684, mIoU - 0.586; best α=0.3), it does not achieve the performance of label smoothing in our scenario.

TABLE III: VAL(p)-dataset scores

Experiment	*Occ*	*mIoU*	*Experiment*	*Occ*	*mIoU*
(C) Base	.630	.566	(F) Stitching (context)	.639	.564
(D) Flipping	.645	.558	(G) Tversky + Focal	.678	.582
(D) Flip+Contrast	.640	.561	(H) Lbl smoothing	.682	.584

I. *Scores on public validation subset*

For comparison, we also provide scores for the publicly available *VAL(p)* subset in Tbl. III. Although the values are very similar, the model can not benefit from the stitching context since the stitching edges in the *VAL(p)* set are not that severe. Example images for the detection on the dataset are presented in Fig. 6.

J. *Scores on test subset*

The best-performing methods on the validation dataset have also been applied to the test set. Although the characteristics of the distress and object instances differ significantly from the validation data, label smoothing performs equally well and increases detection performance considerably (see Tbl. IV). The direct input of stitching maps into the context layer also has a positive effect and reduces false positive detections of patches, thus boosting the occurrence scores of these classes. However, the suggested modification of the loss function does not significantly affect the test data score.

VI. Comparison with Human Labeling

To put the achieved results into perspective, we compared the best result of the automatic distress detection to three human experts in the field. A further evaluation dataset consisting of 13 4x3-meter images (3600 x 2850 pixels) has been created. The dataset was labeled independently by all three experts for the 23 classes. Hence, with the output of our model, there are four different classification results per

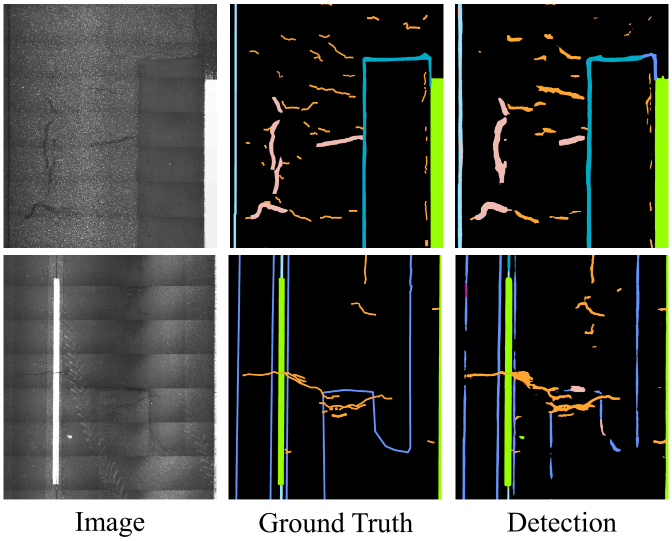


Fig. 6: Exemplary detections on *VAL(p)* (top) and test dataset (bottom). Image are reduced to half of the original image height.

TABLE IV: Class based comparison of occurrence scores. Classes with less than 10 instances are colored in gray.

Exp.	IPCH	APCH	S-Cr	Cr	OJT	PoHL	RVL	SCR	BLE	MRK	WDr	MHL	EXJ	CRb	COB	DLH	MOB	OBJ	JNT	VRG	VEG	IND	NRM	∅ Img
												Validation												
(C) Base	.498	.257	.493	.426	.263	.281	.425	.733	.740	.963	.658	-	.815	.602	.699	.815	.407	.815	.351	.201	.658	-	.989	.610
(D) Aug	.524	.207	.491	.393	.304	**.316**	.499	.733	.443	.960	.749	-	.815	.638	.686	.815	.468	.815	.386	.213	.707	-	.989	.634
(E) ResNet50	**.526**	.381	.447	.372	.188	.235	.254	.776	.569	.955	.505	-	.815	.646	**.743**	.815	.618	.815	.381	.273	.758	-	.989	.633
(F) St-Cntxt	.482	.251	.514	.509	.287	.272	.407	.798	.695	.959	.505	-	.815	.652	.705	.814	.409	.815	**.419**	.251	.648	-	.989	.662
(G) Tv+Fo	.429	.250	**.549**	.524	.223	.158	.433	.733	.631	.963	.646	-	.815	.646	.623	.543	.414	.815	.368	.236	.752	-	.989	.646
(H) L-Smooth	.513	.391	.535	**.595**	.308	.169	.404	.776	.745	**.966**	.646	-	.815	.724	**.707**	.407	.554	.815	.414	**.276**	**.759**	-	.989	**.696**
												Test												
(D) Aug	.213	.469	.573	.362	**.103**	.234	.282	.555	.714	.880	.936	.346	.838	.590	.562	.610	.443	.922	.247	.712	.438	.0	.981	.513
(F) St-Cntxt	**.401**	**.650**	.415	.381	.075	.224	.305	.415	.690	.896	.882	.700	.837	.615	.592	.221	.458	.825	.251	.712	.504	.221	.979	.523
(G) Tv+Fo	.180	.446	.388	.367	.040	.231	.290	.325	.734	**.911**	.882	.700	.885	.651	.602	.480	.377	.886	.155	.710	.506	.221	.982	.516
(H) L-Smooth	.185	.491	**.640**	**.609**	.064	**.317**	**.316**	.263	.733	.900	.936	.787	.885	**.881**	**.850**	.376	.379	.944	**.318**	**.740**	**.534**	.561	.983	**.585**

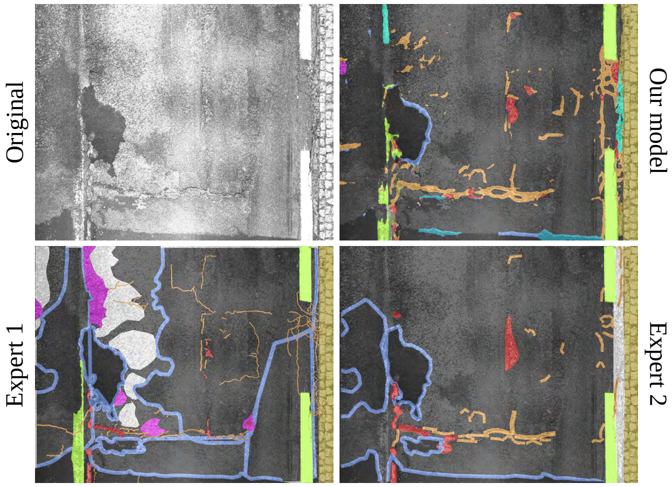


Fig. 7: Comparison between trained experts and the classification result produced by our model. The labels of the experts and the output of our model were rendered in the colors of Fig. 2 and then overlaid upon the input with 50% transparency.

image. An example comparison of a tough image is shown in Fig. 7. The image shows that a precise segmentation of the road classes is a challenging undertaking, even for experts, and that experts also disagree about where different classes begin or end. Furthermore, it can be seen that our model has predicted a classification comparable to the experts.

In addition to a visual evaluation, the three experts and our model have been compared against each other using the mIoU and our Occurrence Score. Therefore, one of the three experts was regarded as ground truth and compared to the results of the other experts and the network output. The process has been repeated three times for each expert. The results of this experiment are given in Fig. 8. The network always performs at least better than one expert in every comparison and, therefore, performs similarly to the trained experts. The experiment also reveals the difference between the labels of the experts, which again highlights the difficulty of the problem at hand.

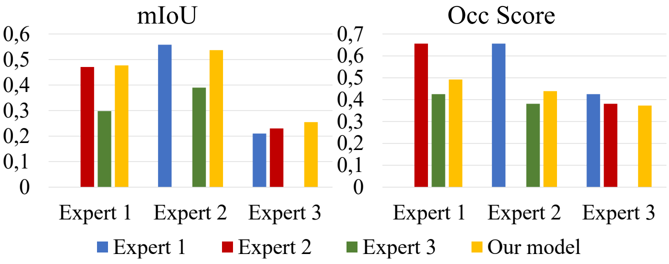


Fig. 8: Labeling results of three human experts and our model compared to each other. The results are divided into three groups, with each expert once representing the ground truth against which all others must compare.

VII. Discussion

We introduced a novel standardized distress catalog describing 23 different distress and object classes essential for road assessment in Austria, Switzerland, and Germany. Furthermore, we have published an evaluation dataset for pixel-perfect segmentation, which includes these classes. We have examined various deep-learning methods and techniques on our datasets and have identified the following as the most important for road segmentation. First, we have shown that classic image augmentation techniques can even hamper classification results because of the well-controlled image acquisition conditions. Second, many classes are challenging to classify without proper context, especially if image artifacts like stitching edges are present in the images. Therefore, integrating context information about these artifacts into the model can significantly improve classification performance. Finally, our research on the labels of various experts has shown that experts struggle and often disagree on the exact segmentation of classes. This also seems to be a problem in training, as it implies that the labels in the training data also contradict each other. Therefore, the results improved significantly when using label smoothing to counteract this problem.

References

[1] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, and H.-M. Gross, "How to get pavement distress detection ready for deep learning? a systematic approach." in *Int. Joint Conf. on Neural Networks (IJCNN)*, 2017, pp. 2039–2047.

[2] R. Stricker, M. Eisenbach, M. Sesselmann, K. Debes, and H. M. Gross, "Improving visual road condition assessment by extensive experiments on the extended gaps dataset," in *Int. Joint Conf. on Neural Networks (IJCNN), Budapest, Hungary*, no. paper N-20496. IEEE, 2019, p. 8.

[3] D. Seichter, M. Eisenbach, R. Stricker, and H. M. Gross, “How to improve deep learning based pavement distress detection while minimizing human effort,” in *IEEE Int. Conf. on Automation Science and Engineering (CASE), Munich*. IEEE, 2018, pp. 63–68.

[4] A. Riid, R. Lõuk, R. Pihlak, A. Tepljakov, and K. Vassiljeva, “Pavement distress detection with deep learning using the orthoframes acquired by a mobile mapping system,” *Applied Sciences*, vol. 9, no. 22, 2019.

[5] R. Augustauskas and A. Lipnickas, “Improved Pixel-Level Pavement-Defect Segmentation Using a Deep Autoencoder,” *Sensors*, vol. 20, 2020.

[6] M. S. Minhas and J. Zelek, “Defect detection using deep learning from minimal annotations,” *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, pp. 506–513, 2020.

[7] H. Lee, “Application of machine vision techniques for the evaluation of highway pavements in unstructured environments,” in *Proc. Fifth International Conference on Advanced Robotics’ Robots in Unstructured Environments*, 1991, pp. 1425–1428.

[8] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, “CrackTree: Automatic crack detection from pavement images,” *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227–238, 2012.

[9] L. Peng, W. Chao, L. Shuangmiao, and F. Baocai, “Research on Crack Detection Method of Airport Runway Based on Twice-Threshold Segmentation,” in *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*. IEEE, 2015, pp. 1716–1720.

[10] H. Oliveira and P. L. Correia, “CrackIT - An image processing toolbox for crack detection and characterization,” in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 798–802.

[11] K. Fernandes and L. Ciobanu, “Pavement pathologies classification using graph-based features,” *2014 IEEE International Conference on Image Processing, ICIP 2014*, pp. 793–797, 2014.

[12] W. Kaddah, M. Elbouz, Y. Ouerhani, V. Baltazart, M. Desthieux, and A. Alfalou, “Optimized minimal path selection (omps) method for automatic and unsupervised crack segmentation within two-dimensional pavement images,” *The Visual Computer*, pp. 1–17, 2018.

[13] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, “Road crack detection using deep convolutional neural network,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3708–3712.

[14] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, “Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection,” *Construction and Building Materials*, vol. 157, pp. 322–330, 2017.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[17] O. Ronneberger, P.Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[18] Z. Wu, T. Lu, Y. Zhang, B. Wang, and X. Zhao, “Crack Detecting by Recursive Attention U-Net,” in *3rd International Conference on Robotics, Control and Automation Engineering*, 2020, pp. 103–107.

[19] J. Huyan, W. Li, S. Tighe, Z. Xu, and J. Zhai, “CrackU-net: A novel deep convolutional neural network for pixelwise pavement crack detection,” *Structural Control and Health Monitoring*, vol. 27, no. 8, 2020.

[20] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, “Deepcrack: Learning hierarchical convolutional features for crack detection,” *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1498–1512, 2019.

[21] R. Lõuk, A. Riid, R. Pihlak, and A. Tepljakov, “Pavement Defect Segmentation in Orthoframes with a Pipeline of Three Convolutional Neural Networks,” *Algorithms*, vol. 13, no. 8, p. 198, 2020.

[22] H. Majidifard, Y. Adu-Gyamfi, and W. Buttlar, “Deep machine learning approach to develop a new asphalt pavement condition index,” *Construction and Building Materials*, vol. 247, 2020.

[24] S. Anand, S. Gupta, V. Darbari, and S. Kohli, “Crack-pot: Autonomous road crack and pothole detection,” *arXiv preprint:1810.05107*, 2018.

[23] S. Chatterjee, A. B. Brendel, and S. Lichtenberg, “Smart infrastructure monitoring: Development of a decision support system for vision-based road crack detection,” in *Proceedings of the International Conference on Information Systems (ICIS), San Francisco, CA, USA*, 2018.

[25] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, “Road damage detection and classification using deep neural networks with smartphone images,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 12, pp. 1127–1141, 2018.

[26] F. Yang, “Feature pyramid and hierarchical boosting network for pavement crack detection,” 2019.

[27] S. Chambon and J. M. Moliard, “Automatic road pavement assessment with image processing: Review and comparison,” *International Journal of Geophysics*, vol. 2011, 2011.

[28] K. Ma, M. Hoai, and D. Samaras, “Large-scale continual road inspection: Visual infrastructure assessment in the wild,” in *Proceedings of British Machine Vision Conference*, 2017.

[29] H. Maeda, T. Kashiyama, Y. Sekimoto, T. Seto, and H. Omata, “Generative adversarial network for road damage detection,” *Computer-Aided Civil and Infrastructure Engineering*, 2020.

[30] Forschungsgesellschaft für Straßen- und Verkehrswesen, *ZTV ZEB-StB - Zusätzliche Technische Vertragsbedingungen und Richtlinien zur Zustandserfassung und -bewertung von Straßen [FGSV-Nr. 489]*. FGSV Verlag, 2006.

[31] RVS 13.01.11, “Quality assurance for structural maintenance, pavement distress catalogue for flexible and rigid pavements,” Vienna, 2009.

[32] RVS 13.01.16, “Quality assurance for structural maintenance, assessment of surface defects and cracks on asphalt and concrete roads,” Vienna, 2013.

[33] P. Maurer, M. Meissner, M. Fuchs, J. Gruber, and P. Foissner, *Straßenzustandserfassung mit dem RoadSTAR – Messsystem und Genauigkeit*. Wien: ÖFPZ Arsenal GmbH, 2002.

[34] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, vol. January, pp. 636–644, 2017.

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[39] S. Tilborghs, T. Dresselears, P. Claus, J. Bogaert, and F. Maes, “3D Left Ventricular Segmentation from 2D Cardiac MR Images Using Spatial Context,” in *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*. Springer, 2020, pp. 90–99.

[40] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *CoRR, abs/1606.02147*, 2016.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[42] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3d fully convolutional deep networks,” in *Machine Learning in Medical Imaging (MLMI)*. Springer International Publishing, 2017.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[44] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada*, 2018.