WildDash - Creating Hazard-Aware Benchmarks

Oliver Zendel(✉), Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernández Domínguez

AIT, Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna, Austria
{oliver.zendel,katrin.honauer.fl,markus.murschitz,daniel.steininger,
gustavo.fernandez}@ait.ac.at

Abstract. Test datasets should contain many different challenging aspects so that the robustness and real-world applicability of algorithms can be assessed. In this work, we present a new test dataset for semantic and instance segmentation for the automotive domain. We have conducted a thorough risk analysis to identify situations and aspects that can reduce the output performance for these tasks. Based on this analysis we have designed our new dataset. Meta-information is supplied to mark which individual visual hazards are present in each test case. Furthermore, a new benchmark evaluation method is presented that uses the meta-information to calculate the robustness of a given algorithm with respect to the individual hazards. We show how this new approach allows for a more expressive characterization of algorithm robustness by comparing three baseline algorithms.

Keywords: Test data · Autonomous driving · Validation · Testing
Safety analysis · Semantic segmentation · Instance segmentation

1 Introduction

Recent advances in machine learning have transformed the way we approach Computer Vision (CV) tasks. Focus has shifted from algorithm design towards network architectures and data engineering. This refers in this context to the creation and selection of suitable datasets for training, validation, and testing.

This work focuses on the creation of validation datasets and their accompanying benchmarks. Our goal is to establish meaningful metrics and evaluations that reflect real-world robustness of the tested algorithms for the CV tasks of semantic segmentation and instance segmentation, especially for autonomous driving (AD). These tasks represent essential steps necessary for scene understanding and have recently seen huge improvements thanks to deep learning approaches. At the same

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01231-1_25) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11210, pp. 407–421, 2018.
https://doi.org/10.1007/978-3-030-01231-1_25

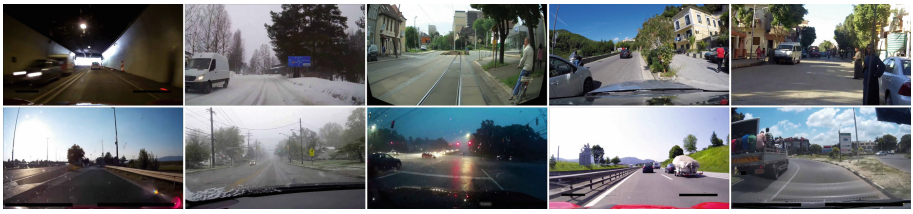

Fig. 1. Examples of hazards found in the *WildDash* dataset. See Table 1 for descriptions.

time, they are basic building blocks of vision-based advanced driver-assistance systems (ADAS) and are therefore employed in high-risk systems.

Demanding CV tasks are becoming increasingly important in safety-relevant ADAS applications. This requires solutions that are robust against many performance-reducing factors (e.g. illumination changes, reflections, distortions, image noise). These factors can be seen as hazards, influences potentially harmful to algorithm performance. Each hazard poses a potential risk and should be tested thoroughly to evaluate the robustness and safety of the accompanying system. Classic risk analysis applied to machine learning systems encompasses an inherent problem: Even if the learning process itself is well-understood, the relation between cause and effect, and the origin of erroneous behaviors are often hard to comprehend: if something goes wrong, it can be difficult to trace back the reason. Incorporating well-categorized test data promises to overcome this issue. Highly expressive meta-information (i.e. describing which aspects and hazards are present in a given test image) allows for reasoning based on empirical evaluations during the test phase: if a statistically significant amount of tests containing a specific hazard fails, it can be assumed that the system is not robust against this hazard. The underlying assumption of this work is: if we use machine-learning-based mechanisms in systems that represent potential risks to human life, a systematic approach comprehensible to humans for testing these components is essential. Only then, sufficient certainty can be obtained regarding the underlying risk and its propagation from one sub-system to others. Data, metrics, and methodologies presented in this work are designed based on this assumption.

Another influential factor regarding the quality of a test set is the inherent dataset bias (see [1]). Most of the publicly available datasets for semantic and instance segmentation in the ADAS context published in recent years still suffer from being too focused on a certain geographical region. These datasets have a strong bias towards Western countries, especially Central Europe. The dataset presented in this work aims to minimize this shortcoming. It embraces the global diversity of traffic situations by including test cases from all over the world. Furthermore, a great variety of different ego vehicles with varying camera setups extracted from dashcam video material is provided. This ultimately results in a vivid cross-section of traffic scenarios, hence the title *WildDash*.

Fig. 2. Example frames of existing datasets. From left to right: CamVid, Cityscapes, KITTI, Playing for Benchmarks, and Mapillary Vistas.

The main contribution of this work is a novel dataset for semantic and instance segmentation, that (i) allows for backtracking of failed tests to visual risk factors and therefore pinpointing weaknesses, (ii) adds negative test cases to avoid false positives, and (iii) has low regional bias and low camera setup bias due to its wide range of sources.

Section 2 gives a thorough overview of existing datasets for semantic and instance segmentation focused on ADAS applications. Section 3 summarizes our process of applying an established risk-analysis method to create a checklist of critical aspects that should be covered by test data to evaluate algorithm robustness. Section 4 explains how we applied the generated checklist and designed our new test dataset: *WildDash*. In Sect. 5, we demonstrate how the additional meta-information about included hazards can be used to create new hazard-aware metrics for performance evaluation. Section 6 describes the training setup of our baseline models and presents detailed segmentation results on specific aspects of *WildDash*. Section 7 gives a short outlook, followed by a summary in Sect. 8.

2 Related Work

2.1 Segmentation Datasets

Brostow et al. [2] introduced *CamVid*, one of the first datasets focusing on semantic segmentation for driving scenarios (see Fig. 2). It is composed of five video sequences captured in Cambridge consisting of 701 densely annotated images, distinguishing between 31 semantic classes. In 2013 the 6D Vision group [3] published the initial version of the *Daimler Urban Dataset* [4]. It contains 5000 coarsely labeled images (*ground*, *sky*, *building*, *vehicle*, *pedestrian*) extracted from two videos recorded in Germany.

The release of the *Cityscapes* Dataset [5] in 2015 marks a breakthrough in semantic scene understanding. Several video sequences were captured in cities across Germany and Switzerland and 25000 images labeled (5000 fine/20000 coarse) with 30 different classes. The corresponding benchmark is still the most commonly used reference, currently listing 106 algorithms for semantic segmentation and 29 algorithms for instance segmentation (July 2018). In the year 2017, the *Raincouver* dataset [6] contributed additional frames depicting road layouts and traffic participants under varying weather and lighting conditions. Published in the same year, *Mighty AI Sample Data* [7] is composed of dashcam images representing different driving scenarios in the metropolitan area of Seattle. The

year 2018 marked two more major contributions in terms of quality and data variability, which represent a further step towards reducing dataset bias. One of them is *Mapillary Vistas Dataset* [8] which contains more than 25000 high-resolution images covering around 64 semantic classes, including varying lighting conditions, locations and camera setups. *Berkeley Deep Drive* [9], on the other hand, specializes more on challenging weather conditions and different times of the day. The *KITTI Vision Benchmark Suite*, first introduced by Geiger et al. [10] in 2012 and aimed at multiple tasks such as stereo, object detection, and tracking was updated in 2018 with ground truth for semantic segmentation [11].

In addition to annotations of real images, a number of synthetically generated datasets emerged in recent years. One of the first contributions to the area of Urban Scene Understanding was *Virtual KITTI* by Gaidon et al. [12] in 2016. It represents a virtual reconstruction of the original KITTI dataset, enhanced by a higher variety of weather conditions. Published in the same year, *SYNTHIA* [13] focuses on multiple scenarios (cities, motorways and green areas) in diverse illumination, weather conditions, and varying seasons. A recent update called *SYNTHIA-SF* [14] furthermore follows the Cityscapes labeling policy. In the following year, Richter et al. [15] introduced the synthetic benchmark suite *Playing for Benchmarks*. It covers multiple vision tasks such as semantic segmentation, optical flow, and object tracking. High-resolution image sequences for a driving distance of 184 Km are provided with corresponding ground-truth annotations.

2.2 Risk Analysis in Computer Vision

A number of publications regarding risk analysis in CV have been published during the last years, since the community seemingly gained awareness for the necessity to train and test for increasingly difficult conditions.

In 2015, Zendel et al. [16] introduced the concept of risk analysis for CV tasks. In contrast to high-level driving hazards (e.g. car crash, near-miss events as in the SHRP 2 NDS database [17]), this work focuses on visual hazards (e.g. blur, glare, and overexposure). They create a checklist of such hazards that can impair algorithm performance. The list has more than 1000 generic entries which can be used as seeds for creating specialized entries for individual CV tasks. Such were presented for stereo vision in 2017 in *Analyzing Computer Vision Data* [18] where they strongly emphasize on the underrated aspect of *negative test cases*. These are tests where algorithms are expected to fail. Since most of the data is highly focused on training, many works do not consider the negative test class, neither in the evaluation metric nor in the data itself. For a safe and robust system it is important that an algorithm does not 'overreact' and knows when it is not able to provide a reliable result. No indications have been found in any of the mentioned evaluation frameworks and benchmarks that true negative test cases are evaluated. Most common is the *don't-care*-approach (e.g. in Cityscapes), where all the regions that are annotated using a negative (=unknown/invalid) class are not evaluated. This means that an image containing only negative classes is not evaluated at all.

Both risk analysis publications [16] and [18] include interesting claims and tools for measuring and improving test data quality. However, the authors only apply their concepts to existing test datasets and do not create a new dataset themselves.

In this work we are trying to build upon their work and actually create a dataset allowing for hazard-aware evaluation of algorithms. In addition, *WildDash* deliberately introduces negative test cases to close this crucial gap.

3 Risk Analysis

The process of collecting a comprehensive list of factors that pose risks to a system and the overall assessment of these risk factors is called risk analysis. For the course of the *WildDash* dataset, we started with the results from a publicly available generic CV risk analysis called CV-HAZOP [16]. The generic entries from this list are *concretized* to create a version specific to the current task at hand. The first step of conducting the risk analysis is the definition of the CV task itself that shall be evaluated.

We designed our dataset as an organic extension to existing datasets. Thus, we chose to use a task definition close to the one used in the popular Cityscapes [5] dataset. It provides a valuable tool solving important tasks for autonomous driving: navigation, scene understanding and collision avoidance. The task definition categorizes test cases: those which are in-scope as *positive* test cases vs. those lying outside the task definition as *negative* test cases.

3.1 Task Definition: Semantic Segmentation

The algorithm shall assign a single best fitting label to each pixel of a given color image. The specific labels and semantics for these labels can be found in Cordts et al. [5] and focus on scene understanding for autonomous driving.

In essence, the task focuses on assigning each pixel in an image to exactly one of these possible classes: *road*, *sidewalk*, *parking*, *rail track*, *person*, *rider*, *car*, *truck*, *bus*, *on rails*, *motorcycle*, *bicycle*, *caravan*, *building*, *wall*, *fence*, *guard rail*, *bridge*, *tunnel*, *pole*, *traffic sign*, *traffic light*, *vegetation*, *terrain*, *sky*, *ground*, *dynamic*, and *static*.

All scenes depict frontal vehicle views of traffic scenarios. The camera angle and orientation should be comparable to a human driver or co-driver. It can be positioned outside the vehicle or behind the windscreen.

Some of the labels do not affect the results because they are not part of the evaluation in the Cityscapes benchmark. Other labels cause varying annotations, as the corresponding concepts are hard to narrow down into a concrete task description for an annotator. To correct this, we deviate from the original work of Cordts et al. [5] as follows:

- The *trailer* label is not used. Trailers are labeled as the vehicle that is attached to it and parked trailers without an attached vehicle as *dynamic*.

- The label *pole group* is not used. These parts are labeled as *pole*.
- Areas within large gaps in an instance label are annotated by the content visible in that hole, in contrast to being filled with the enclosing label (original Cityscapes). Whenever content is clearly visible through the hole consisting of more than just a few pixels, it is annotated accordingly.

The original Cityscapes labels are focusing on German cities. We are refining and augmenting some of the definitions to clarify their meaning within a broader worldwide context:

- Construction work vehicles and agriculture vehicles are labeled as *truck*.
- Overhead bridges and their support pillars/beams are labeled as *bridge*. Roads/sidewalks/etc. on bridges still keep their respective labels.
- Two/Three/Four-wheeled muscle-powered vehicles are labeled as *bicycle*.
- Three-wheeled motorized vehicles are labeled as *motorcycle* (e.g. auto rickshaws, tuk-tuk, taxi rickshaws) with the exception of vehicles that are intended primarily for transport purposes which get the *truck* label.

3.2 Task Definition: Instance Segmentation

Instance segmentation starts with the same task description as semantic segmentation but enforces unique instance labels for individual objects (separate labels even for adjoint instances). To keep this benchmark compatible with Cityscapes, we also limit instance segmentation to these classes: *person*, *rider*, *car*, *truck*, *bus*, *on rails*, *motorcycle*, *bicycle*, *caravan*.

3.3 Concretization of the CV-HAZOP List

The concretization process as described in *Analyzing Computer Vision Data* [18] starts from the generic CV-HAZOP list. Using the task definitions (Sects. 3.1 and 3.2), the relevant hazards are filtered. In our case, we filtered out most temporal effects (as the task description requires a working algorithm from just one image without other sequence information). The remaining entries of the list were reviewed and each fitting entry was reformulated to clearly state the hazard for the given task definition.

3.4 Clustering of Hazards

Getting a specific evaluation for each identified hazard would be the ideal outcome of a hazard-aware dataset. However, real-world data sources do not always yield enough test cases to conclusively evaluate each risk by itself. Furthermore, the effects seen within an image often cannot be attributed to a single specific cause (e.g. blur could either be the result of motion or a defocused camera). Thus multiple risks with common effects on output quality were clustered into groups. The concretized entries have been clustered into these ten risk clusters: blur, coverage, distortion, hood, occlusion, overexposure, particles, underexposure, variations, and windscreen. See Table 1 for an explanation of each risk cluster and Fig. 1 for example images containing these hazards.

Table 1. Risk clusters for *WildDash*. Figure 1 contains examples in the same order

Risk cluster	Hazard examples
blur	Effects of motion blur, camera focus blur, and compression artifacts
coverage	Numerous types of road coverage and changes to road appearance
distortion	Lens distortion effects (e.g. wide angle)
hood	Ego-vehicle's engine cover (bonnet) is visible
occlusion	Occlusion by another object or the image border
overexposure	Overexposed areas, glare and halo effects
particles	Particles reducing visibility (e.g. mist, fog, rain, snow)
underexposure	Underexposed areas, twilight, night shots
variations	Intra-class variations, uncommon object representations
windscreen	Windscreen smudges, raindrops and reflections of the interior

4 WildDash Setup

4.1 Dataset Collection

Gathering a lot of challenging data without strong content bias is a hard task. Therefore, the input images of our dataset are collected from contributions of many 'YouTube' authors who either released their content under CC-BY license or individually agreed to let us extract sample frames from their videos. Potential online material is considered of interest with regard to the task descriptions (Sects. 3.1 and 3.2) if it met the following requirements: (i) data was recorded using a dashcam, (ii) front driving direction, (iii) at least one hazard situation arises, (iv) some frames before and after the hazard situation exist. This allows for a later expansion of our dataset towards semantic flow algorithms. All such videos are marked as a potential candidate for *WildDash*. From the set of candidate sequences, individual interesting frames were selected with the specific hazards in mind. Additionally, the content bias was reduced by trying to create a mixture of different countries, road geometries, driving situations, and seasons.

This selection resulted in a subset of about 1800 frames. A meta-analysis was conducted for each frame to select the final list of frames for the public validation and the private benchmarking dataset.

4.2 Meta-data Analysis

In order to calculate hazard-aware metrics the presence of hazards in each frame needs to be identified. Another design goal of *WildDash* is limited redundancy and maximal variability in domain-related aspects. Therefore, (i) domain-related and (ii) hazard-related meta-data is added to each frame. The following predefined values (denoted as set {.}) are possible:

- Domain-related: *environment* {'city', 'highway', 'off-road', 'overland', 'suburban', 'tunnel', 'other'} and *road-geometry* {'straight', 'curve', 'roundabout', 'intersection', 'other'}.

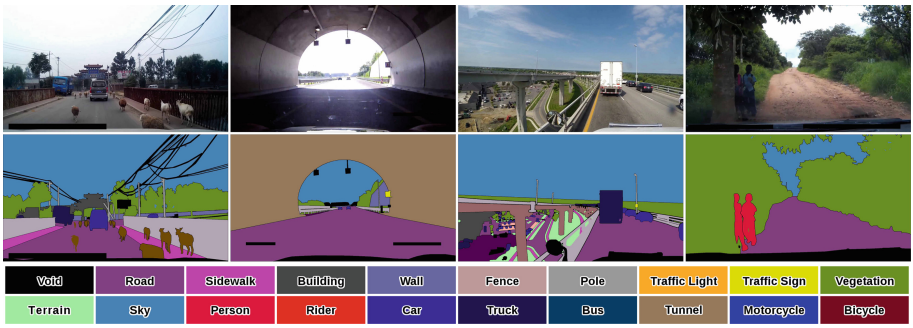

Fig. 3. Positive test cases from wd_val_01 (cn0000, si0005, us0006, and zm0001) together with a visualization of the respective semantic segmentation and color legend.

- Hazards-related: One severity value {'none', 'low', 'high'} for each of the ten risk clusters from Table 1.

The severity for a given risk is set to 'high' if large parts of the image are clearly affected or the appearance of humans/vehicles is affected. All other occurrences of the risk are represented by 'low' severity or if not present by 'none'.

4.3 Positive Test Cases

Based on the meta list, a diverse set of test frames covering each of the hazards has been selected and separated into a public validation set (wd_val_01, GT is published) of 70 test cases and a hidden benchmark set (wd_bench_01, GT is withheld) of 141 test cases. The GT has been generated using a dedicated annotation service and many additional hours by the authors to ensure consistent quality. Figure 3 shows a few examples taken from the WildDash public validation set.

4.4 Negative Test Cases

One of the central requirements presented by Zendel et al. [18] is the inclusion of negative test cases: tests that are expected to fail. The point of having these images in the dataset is to see how the system behaves when it is operating outside its specifications. A robust solution will recognize that it cannot operate in the given situation and reduce the confidence. Ideally, a perfect system flags truly unknown data as invalid. Table 2 lists test cases which increasingly divert from the region of operation of a regular assisted driving system while Fig. 4 shows some of the respective input images. With 141 positive and 15 negative test cases the WildDash benchmarking set wd_bench_01 contains a total of 156 test cases.

Table 2. Negative test cases from wd_bench_01.

Altered valid scenes		Abstract/Image noise	
wd0141	RGB/BGR channels switched	wd0142	White wall close-up
wd0143	Black-and-white image	wd0144	Digital image receive noise
wd0148	Upside-down version	wd0146	Analog image receive noise
wd0151	Color-inverted image	wd0147	Black image with error text
wd0155	Image cut and rearranged	wd0154	Black sensor noise
Out-of-scope images			
wd0145	Only sky with clouds		
wd0149	Macro-shot anthill		
wd0150	Indoor group photo		
wd0152	Aquarium		
wd0153	Abstract road scene with toys		

Fig. 4. Negative test cases wd0141, wd0142, wd0145, wd0146, and wd0152. See Table 2 for content descriptions

5 Hazard-Aware Evaluation Metrics

The meta-analysis of the dataset allows for the creation of subsets for each of the identified hazard clusters. For each group, all frames are divided by severity into three groups: none, low and high. Performance evaluation can be conducted for each severity-subset to obtain a coarse measure of the individual hazard's impact on an algorithm's performance. The Intersection over Union (IoU) measure [19] represents the 'de facto' established metric for assessing the quality of semantic segmentation algorithms. For each label the ratio of true positives (i.e. the intersection of predicted and annotated labels) over the union of true positives, false positives and false negatives is evaluated. The IoU scores per label class are averaged to calculate a single performance score per hazard subset called mean IoU (mIoU). The *impact* of the individual hazard reflects its negative effect on the algorithm's performance. It is calculated as: $r_{impact} = 1.0 - \frac{min(mIoU_{low}, mIoU_{none})}{max(mIoU_{low}, mIoU_{high})}$. Therefore, a value of 0.0 implies no impact, while a score of e.g. 0.5 corresponds to a hazard of reducing performance by 50%. The subset *low* represents border cases between influential and non-influential test cases and thus $mIoU_{low}$ is present at both numerator and denominator.

Occlusions are only relevant for foreground objects with instance annotations. To mitigate this, the risk cluster *occlusions* evaluates only labels with instance

annotations (human and vehicle category) and ignores the single label with the largest area (as this is normally the fully visible occluder).

5.1 Evaluating Negative Test Cases

Evaluation of negative test cases might seem straight forward at first: per definition we expect an algorithm to fail for negative test cases in a graceful manner, i.e. mark the output as invalid. This creates a paradox situation: output marked as invalid is considered to be correct while any other output is counted as incorrect. This binary form of evaluation is not very appropriate, especially as the borderline between positive and negative test cases is ambiguous. Just because a specific situation/aspect is not clearly stated in the domain/task definition does not make it a clean negative test case (i.e. 'algorithm must fail here'). Often, a test case states a situation that is clearly not part of the system's task definition; for example, an upside down image of a street scene. It is still possible to assign unambiguous legitimate semantic labels for this test image. In these cases, we treat all algorithm output as correct, that is either equal to such legitimate label, or marked as invalid.

6 Evaluation

This section provides first valuable insights concerning opportunities and shortcomings of recently published datasets predominantly used in the research field of semantic segmentation. For this purpose, three baseline models (i.e. cityscapes, mapillary, mapillary+) varying with regard to the amount and source of training data, were trained from scratch and thoroughly evaluated on subsets of the *WildDash* dataset representing specific visual hazards.

6.1 Experimental Setup

This section describes the setup of the baseline models, which are based on the pytorch implementation of Dilated Residual Networks (drn) [20]. Employing dilated convolution for semantic segmentation facilitates an efficient aggregation of features at multiple scale levels without losses introduced by downsampling. To ensure comparability between all models, each experiment has been carried out with the same training configuration. The network architecture drn-d38 was selected due to the balance between labeling accuracy and training duration it provides. Moreover, the input batches consist of 8 pairs of input images and corresponding annotations each, and are randomly rescaled by a factor between 0.5 and 2 to improve scale invariance, randomly flipped in horizontal direction, and finally randomly cropped to a size of 896×896 pixels. As a pre-processing step, the *Mapillary Vistas* dataset has been rescaled and cropped to fit the resolution of Cityscapes (2048×1024 pixels). Since the *Cityscapes* dataset consists of 3475 pixel-level annotations, subdivided into 2975 training and 500 validation images, and therefore provides the least amount of training data, a subset of

Table 3. mIoU scores of the conducted experiments on varying target datasets

Baseline model/dataset	Cityscapes	Mapillary	WildDash (val/bench)	WildDash negative test cases
cityscapes	**63.79**	30.31	16.5/15.4	7.2
mapillary	44.81	50.24	29.3/27.4	12.9
mapillary+	46.34	**52.34**	**30.7/29.8**	**27.4**

Mapillary with a similar number of images has been used to train the comparable baseline method, further referred to as mapillary. During our experiments the 1525 *Cityscapes* and 5000 *Mapillary* test images are not included, since they are withheld for benchmarking purposes and thus not publicly available. The baseline method mapillary+ uses all publicly available *Mapillary* data of 18000 training and 2000 validation images. To cope with the increased amount of sampled input data a faster decay of the learning rate was achieved by lowering the step size from 100 to 17 epochs during the last experiment. Training input has been restricted to the labels evaluated in the *WildDash* benchmark without performing any further label aggregation.

6.2 Cross-Dataset Validation

To quantify shortcomings and the degree of variability inherent to semantic segmentation datasets, the learned models are validated on three target datasets. A detailed overview of the corresponding evaluation is given in Table 3.

As expected, the models perform best on the datasets they have been trained on. The highest mIoU of 63.79 is achieved by the cityscapes model. However, the validation set of the *Cityscapes* dataset consists of only three image sequences captured in Central European cities. The results of this model on datasets like *Mapillary* and *WildDash* show that training solely on *Cityscapes* images is insufficient to generalize for more challenging ADAS scenarios. The model cannot cope with visual hazards effectively. The highest score on *WildDash* is achieved by the mapillary+ experiment with mIoU scores of 30.7 on validation and 29.8 on the test set, based on more distinct scene diversity and global coverage present within the training data of Mapillary. Exemplary results of our baseline experiments on *WildDash* validation images are shown in Fig. 5. As long as input images bear a high resemblance to the training set of *Cityscapes*, as shown in the first row, no significant loss in labeling performance occures. However, models like mapillary and mapillary+ are clearly more robust to the challenging *WildDash* scenarios.

6.3 Testing Visual Hazards

Detailed results on varying subsets of the *WildDash* test dataset, representing a diverse range of visual hazards, are reported in Table 4[1]. As expected, the influence of the individual hazards is clearly reflected in the algorithm performance.

[1] See supplementary material for additional results including instance segmentations.

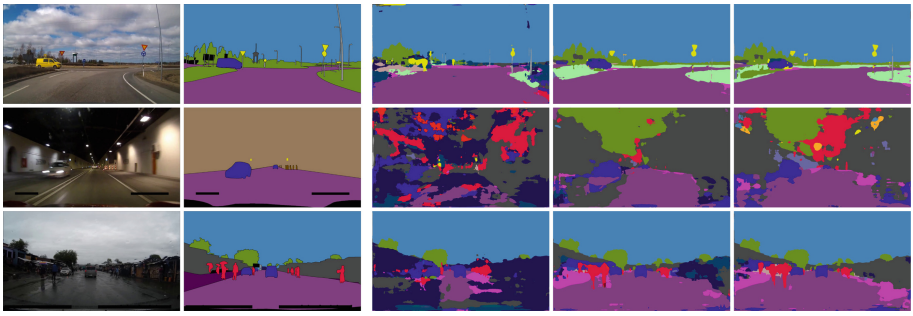

Fig. 5. Qualitative results of our baseline models on *WildDash* validation images (left to right: input image, corresponding ground truth, and the inferred labelings of our baseline models cityscapes, mapillary, and mapillary+)

Table 4. mIoU scores of the baseline model mapillary+ on hazard-related *WildDash* subsets, grouped by their severity of the respective hazard. The impact score, which is introduced in Sect. 5, quantifies the potential negative influence of a specific hazard on the labeling performance

hazard	blur	coverage	distortions	hood	occlusion	overexp	underexp	particles	windscreen	variations
none	29.0	31.0	**31.4**	**32.9**	26.4	**32.2**	**31.5**	**30.2**	**31.8**	29.0
low	**32.2**	28.6	28.2	27.8	**32.1**	23.5	31.0	29.3	28.5	**30.7**
high	26.6	**32.8**	26.8	22.4	30.4	17.0	20.8	29.3	27.8	27.9
impact	0.17	0.08	0.15	0.32	0.05	**0.47**	0.34	0.03	0.12	0.09

Evaluating hazards causing significant image degradations (e.g. blur, over- and underexposure) show an high impact, thus leading to lower algorithm performance. On the other hand, effects caused by lens distortions lead to a graceful decrease of labeling accuracy. Furthermore, mixing environmental effects such as fog and heavy rain with slight snowfall, leads to high variations in algorithm performance. This will be considered in the future, by partitioning the risk cluster *particles* as two disjunct subsets.

6.4 Testing Domain-Related Aspects

As already discussed, another important aspect of test data is a distinctive and comprehensive coverage of domain aspects, such as differences regarding environments and varying types of road layouts. The influence of these aspects is presented in Table 5. As the results show, labeling performance varies strongly with regard to the domain. Unsurprisingly, tunnel scenes tend to yield inferior accuracy due to a mixture of low light conditions and homogeneously textured regions, as well as their relatively rare occurrence within the training data. The algorithm performs robust in the city, sub-urban, and overland domain, which can be explained by the high number of learned urban scenes, constituting 90 percent of the *Mapillary* dataset and the low complexity of overland scenes. As

Table 5. mIoU scores of the baseline model mapillary+ on domain subsets of *WildDash*

domain	city	highway	offroad	overland	suburban	tunnel	curve	intersection	roundabout	straight
mIoU	31.3	24.5	**32.7**	29.3	31.6	19.6	28.7	31.7	**36.6**	28.0

for variations in road layouts, the best labeling scores are achieved in roundabout scenes, followed by those containing intersections. This could be caused by the strong uniformity present within these subgroups and lower vehicle speeds leading to reduced motion blur.

6.5 Negative Test Cases

Labeling results of negative test cases show typical characteristics dependent on the specific subgroup. Representative qualitative results are shown in Fig. 6. If the system is confronted with upside-down images, the trained model partially relies on implicitly learned location priors, resulting in a clearly visible labeling conflict between road and sky in the top region. Labeling performance on abstract test cases, on the other hand, is strongly influenced by image noise and high-frequency texture features, leading to a drift towards properties resembling similar labels. The significantly lower confidence scores of altered and out-of-scope images may be used to suppress the labeling partially or completely, giving the system the ability to recognize cases where it is operating outside its specification.

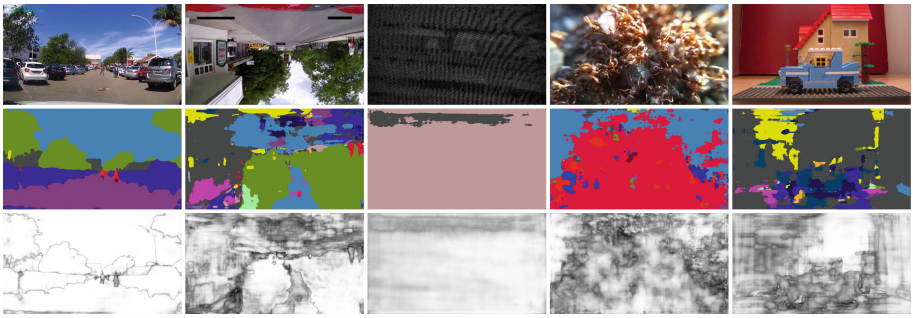

Fig. 6. Input images, semantic segmentation results and corresponding confidence of baseline model mapillary+ on *WildDash* test images (left to right: positive test case, altered valid image, abstract image and two out-of-scope images).

7 Outlook

The benchmark has now started its operation at the website wilddash.cc. It allows everyone to submit their algorithm results for evaluation. In the future, we want to increase the number of validation and benchmark images, as well as

the number of test cases for each hazard cluster (especially for the high severity subsets). Also, the number of hazard clusters will most probably increase. All those improvements and extensions will be adapted according to the results of upcoming submissions. We are confident, that user feedback will help us to improve and advance *WildDash* and the concept of hazard-aware metrics in general.

8 Conclusions

In this paper we presented a new validation and benchmarking dataset for semantic and instance segmentation in autonomous driving: *WildDash*. After analyzing the current state-of-the-art and its shortcomings, we have created *WildDash* with the benefits of: (i) less dataset bias by having a large variety of road scenarios from different countries, roads layouts as well as weather and lighting conditions; (ii) more difficult scenarios with visual hazards and improved meta-information, clarifying for each test image which hazard is covered; (iii) inclusion of negative test cases where we expect the algorithm to fail.

The dataset allows for hazard-aware evaluation of algorithms: The influence of hazards such as blur, underexposure or lens distortion can directly be measured. This helps to pinpoint the best areas for improvements and can guide future algorithm development. Adding negative test cases to the benchmark further improves *WildDash*'s focus on robustness: we look even beyond difficult test cases and check algorithms outside their comfort zone. The evaluation of three baseline models using *WildDash* data shows strong influence of each separate hazard on output performance and therefore confirms its validity. The benchmark is now open and we invite all CV experts dealing with these tasks to evaluate their algorithms by visiting our new website: `wilddash.cc`.

Acknowledgement. The research was supported by ECSEL JU under the H2020 project grant agreement No. 737469 AutoDrive - Advancing fail-aware, fail-safe, and fail-operational electronic components, systems, and architectures for fully automated driving to make future mobility safer, affordable, and end-user acceptable. Special thanks go to all authors who allowed us to use their video material and Hassan Abu Alhaija from HCI for supplying the instance segmentation example algorithms.

References

1. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR, pp. 1521–1528 (2011)
2. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 44–57. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_5
3. Franke, U., Gehrig, S., Rabe, C.: Daimler Böblingen, 6D-Vision. http://www.6d-vision.com. Accessed 15 Nov 2016

4. Scharwächter, T., Enzweiler, M., Franke, U., Roth, S.: Stixmantics: a medium-level model for real-time semantic scene understanding. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 533–548. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_35
5. Cordts, M., et al.: The cityscapes dataset. In: CVPR Workshop on The Future of Datasets in Vision (2015)
6. Tung, F., Chen, J., Meng, L., Little, J.J.: The raincouver scene parsing benchmark for self-driving in adverse weather and at night. IEEE Robot. Autom. Lett. **2**(4), 2188–2193 (2017)
7. Mighty AI: Mighty AI Sample Data. https://info.mty.ai/semantic-segmentation-data. Accessed 07 Mar 2018
8. Mapillary Research: Mapillary Vistas Dataset. https://www.mapillary.com/dataset/vistas. Accessed 16 Feb 2018
9. University of California, Berkeley, U.: Berkeley deep drive. http://data-bdd.berkeley.edu/. Accessed 07 Mar 2018
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR (2012)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: The KITTI Vision Benchmark Suite. http://www.cvlibs.net/datasets/kitti/eval_semantics.php. Accessed 16 Feb 2018
12. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
13. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
14. Hernandez-Juarez, D., et al.: Slanted stixels: representing San Francisco steepest streets. In: BMVC (2017)
15. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV (2017)
16. Zendel, O., Murschitz, M., Humenberger, M., Herzner, W.: CV-HAZOP: introducing test data validation for computer vision. In: ICCV (2015)
17. Transportation Research Board of the National Academy of Sciences: The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset. Available from the SHRP 2 NDS InSight Data Dissemination web site (2013)
18. Zendel, O., Honauer, K., Murschitz, M., Humenberger, M., Dominguez, G.F.: Analyzing computer vision data - the good, the bad and the ugly. In: CVPR, pp. 6670–6680 (2017)
19. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
20. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: CVPR (2017)