# Unifying Panoptic Segmentation for Autonomous Driving

Oliver Zendel Matthias Schörghuber Bernhard Rainer Markus Murschitz
Csaba Beleznai
AIT Austrian Institute of Technology
oliver.zendel,matthias.schoerghuber,bernhard.rainer,markus.murschitz,csaba.beleznai@ait.ac.at

## Abstract

*This paper aims to improve panoptic segmentation for real-world applications in three ways. First, we present a label policy that unifies four of the most popular panoptic segmentation datasets for autonomous driving. We also clean up label confusion by adding the new vehicle labels pickup and van. Full relabeling information for the popular Mapillary Vistas, IDD, and Cityscapes dataset are provided to add these new labels to existing setups.*

*Second, we introduce Wilddash2 (WD2), a new dataset and public benchmark service for panoptic segmentation. The dataset consists of more than 5000 unique driving scenes from all over the world with a focus on visually challenging scenes, such as diverse weather conditions, lighting situations, and camera characteristics. We showcase experimental visual hazard classifiers which help to pre-filter challenging frames during dataset creation.*

*Finally, to characterize the robustness of algorithms in out-of-distribution situations, we introduce hazard-aware and negative testing for panoptic segmentation as well as statistical significance calculations that increase confidence for both concepts. Additionally, we present a novel technique for visualizing panoptic segmentation errors.*

*Our experiments show the negative impact of visual hazards on panoptic segmentation quality. Additional data from the WD2 dataset improves performance for visually challenging scenes and thus robustness in real-world scenarios.*

## 1. Introduction

During the last years, the previously separate tasks of semantic scene segmentation (assigning a semantic label like car, road, street sign to each pixel) and instance segmentation (assigning masks per individual instance) have been combined into the panoptic segmentation task [15].

Diverse challenges imposed by real-world autonomous driving applications confront ML systems with data distributions different from those used during training. Their ability to extrapolate to out-of-distribution (OOD) test cases is an active but largely unsolved problem. The combination of multiple datasets promises a partial solution by combining different advantages and mitigating individual shortcomings. In this paper, we present both a unification method for existing road scene datasets and the new dataset *Wilddash2* based on this principle. Recent work of Hendrycks *et al.* [9] shows that while some robustness-related distribution shifts can be synthetically generated from data, other factors (*e.g.* location/scene-specific image content) can only be well represented during the image formation process of dataset creation. Inspired by this, *Wilddash2* is captured at diverse locations (see Figures 1,2), environment conditions, and includes many potentially performance-reducing factors (called visual hazards [40]) such as: fog, occlusions, overexposure and many more. Additionally, for benchmarking we add many out-of-

Figure 1. Diverse driving scenes from Wilddash2; ae0021: mirroring wet road in UAE, ar0006: broad avenue from Argentia, ci0011: busy market in Côte d'Ivoire, do0007: unusual pickup from Dominican Republic, ee0031: night scene from Estonia with a highly reflective car hood, gr0027: rainy drive in Greece

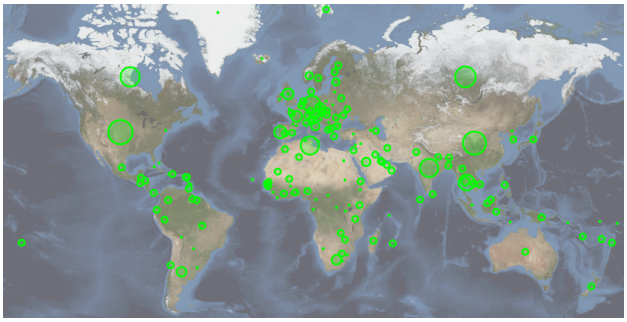

Figure 2. Visualization of Wilddash2 geographic distribution. Dots denote 1-9 scenes; small circles 10-50; medium circles: 50-200; large circles: >200 scenes. Globe courtesy of USGS [35].

domain frames (e.g. a blank frame) to test for false positives called *negative testing*.

The most prominent novelties presented in this paper are: (a) introduction of a unified label policy enclosing and backward compatible to the popular datasets Mapillary Vistas (MVD), Cityscapes, Indian Driving Dataset (IDD), and Wilddash, including two new vehicle labels *pickup* and *van*. (b) a new dataset and benchmark service with a public leaderboard for the panoptic segmentation of driving scenes called *Wilddash2* supporting the unified label policy. (c) methods to improve panoptic segmentation using hazard-awareness, negative testing, supercategories, and a new form of visualizing differences between prediction results and the ground truth (GT). (d) a method to analyze the statistical significance of the calculated visual hazard impact on output performance. (e) panoptic segmentation experiments using Wilddash2 and learned visual hazard classifiers to automatically detect visually challenging situations in camera data.

Section 2 summarizes the current state of the art for panoptic segmentation datasets. Section 3 presents a new public panoptic segmentation dataset. Section 4 introduces multiple tools to improve the evaluation and benchmarking of panoptic segmentation while Section 5 analyses how to calculate the statistical significance of hazard-aware testing. The experimental Section 6 showcases examples of panoptic segmentation using the new dataset and results from classifier experiments to automatically identify visual hazards. All achievements and results are summarized in the final Section 7.

## 2. State-of-the-Art

Solutions for accomplishing real-world vision tasks robustly need to consider the underlying *open world* assumption: no task specification enclosing all potential variations is achievable. This requires establishing datasets with vast diversity, often considering OOD data. Learning unambiguous concepts from ambiguous data needs adequate protocols and metrics to quantify ambiguous image content.

Many datasets have been proposed recently to enhance situational diversity in terms of imaging conditions (e.g. weather, visibility). The Raincouver Scene Parsing benchmark [34], Dark Zurich dataset [31], ADUULM dataset [26], the BDD100K dataset [38], the synthetic FoggyCityscapes [30], and the Woodscape dataset [37] present driving scenes each adding some adverse condition (fog, rain, daytime, dusk, night). Exclusively Dark (ExDark) dataset [19] aims at extending object detection towards low-light situations. The recent Adverse Conditions (ACDC) dataset [32] provides detailed semantic segmentation, images depicting both normal and adverse conditions, and characterizes uncertainties associated with specific viewing conditions. NVIDIA's ClearSightNet [25] (part of NVIDIA DRIVE) calculates per-pixel measures of occlusions and visibility reductions via a lightweight convolutional neural network.

Another prevailing scheme to enhance dataset diversity is the integration of OOD samples. The Lost and Found dataset [27] proposes an OOD-focused dataset (using the Cityscapes dataset [4] as their baseline) and the Fishyscapes Benchmark [1] introduces a public benchmark for semantic segmentation with a special focus on OOD detection. The A2D2 dataset [7] proposes OOD sample detection and similarity-based clustering of OOD samples. The Combined Anomalous Object Segmentation (CAOS) benchmark dataset [8] integrates BDD100K with synthetic OOD object overlays. OOD samples at scene-domain level are targeted in the TAS500 dataset [22] which provides semantic labeling for autonomous driving in unstructured environments. Synthetic data can also be used to enrich the learning process and to extend learned representations beyond common domains. The VIPER [29] dataset and benchmark use scenes from GTA5 as a baseline to create a driving scenes dataset. This allows for the generation of large datasets with low label noise but adds the specific rendering artifacts and digital asset quality as considerable dataset bias. Apolloscapes [11] focuses on sensor fusion and supplies panoptic annotated LiDAR data using a simplified label policy. Panoramic panoptic datasets WildPPS [14], KITTI-360 [17] provide annotations for fisheye-camera data creating full 360°driving scenes

Nowadays, driven by legal authorities and regulatory bodies, the standardization community is aware of the arising importance of scene interpretation in cars (part of situational awareness). The ISO Central Secretary published the guideline ISO/PAS 21448:2019 [13] which specifically addresses the problem of visual hazards (called triggering events), such as overexposure or weather-related effects.

Despite the various adverse-situation-oriented datasets, the scientific community has predominantly adopted four road scene datasets, therefore strongly affecting the scien-

tific evolution of semantic road scene understanding. These datasets offer diversity, dataset scale, and annotations covering the needs of recent vision tasks:

- The Cityscapes dataset [3] in 2016 was the first extensive dataset for scene understanding supplying 5000 scenes with 35 different classes from 50 cities in Central Europe. Its benchmark service is still the most used reference for comparisons and added panoptic segmentation in 2019. Location, lighting conditions, and weather are very uniform and controlled. It uses a license similar to CC-BY-NC 4.0.
- The Mapillary Vistas dataset (MVD) [24], released in 2017, represents a strong increase in size ($20k$ frames with GT), worldwide scope, and 64 labels (40 with instances). It is predominantly focusing on daytime, clear-weather scenarios, and is supplied under a CC-BY-NC-SA 4.0 license.
- The *Wilddash* [39] dataset and benchmark service introduced two concepts to improve characterization of algorithms: Hazard-aware testing and use of negative test cases. It uses the Cityscapes label policy and only supplies around 220 frames for benchmarking and validation under a license similar to CC-BY-NC 4.0.
- The Indian Driving Dataset (IDD) [36] from 2019 supplies 10k frames from Indian cities with very dense and unstructured driving scenarios. Its label policy is largely oriented on the Cityscapes policy but introduces new fall-back classes. Mainly composed of clear-weather daylight footage from only 150 driving sequences[1].

# 3. Dataset Design

We present Wilddash2, a new dataset for robust panoptic segmentation training and evaluation combining the most valuable features of the four previously identified panoptic segmentation datasets.

## 3.1. Frame Selection

The frame selection for Wilddash2 focuses on the same principles as the Wilddash [39] dataset: visually challenging driving scenes from all over the world.

In general, driving datasets consist of scenes limited to a single regional area (*e.g.* Cityscapes: Central Europe, IDD: India). Public dashcam videos from over 150 countries in the world are used to create Wilddash2 reducing this regional dataset bias. This includes more than 2000 frames from historically underrepresented areas such as Africa, Middle Eastern countries, and Oceania. Figure 2 shows a visual representation of the broad geographic spread of WD2 frames.

[1]No clear license text is distributed with IDD; their homepage suggests a CC-BY-NC-like license.

The collection of videos included targeted searches for underrepresented regions and difficult scenes. We manually selected interesting frames and annotated the severity of potentially degrading performance factors as *visual hazards* [39]: blur, road-coverage, lens distortion, hood (visibility of car bonnet), occlusions, underexposure, overexposure, particles (fog, rain, snow), screen (windshield visibility and interior reflections), and variations (rare variations of vehicles and attire). The severity level of each *visual hazard* was qualitatively annotated using *none*, *low* or *high* (see [39]). The top of Table 1 shows the percentage of visual hazards present in the frames of the dataset.

The final list of Wilddash2 frames is selected based on these annotations to provide a balanced mix of identified hazards and domain aspects. To limit redundancy, we ensured that there is no direct visual or contextual overlap between frames in the dataset. In terms of quantity, Wilddash2 is offering 5032 scenes, comparable to Cityscapes's 5000 frames and more than 20 times the amount of Wilddash. The dataset is distributed freely under the CC-BY-NC license. To conform to data protection rules, the access is limited to registered scientific users. This allows WD2 to include all frames in unaltered form to prevent unnecessary training and evaluation bias (*e.g.* training with blurred faces can mislead the network into classifying blurred blobs as faces). Wilddash2 includes a separate version with pseudonymized RGB images for use in publications.

## 3.2. Label policy

We have created a unified label policy for Wilddash2 that merges the labels of MVD, Cityscapes, and IDD. This includes the Wilddash dataset, as its label policy is based entirely on Cityscapes.

Unification involves three operations:

- Union of labels: the union of all base labels from MVD, Cityscapes, and IDD is used as a starting point. Duplicate labels are merged.
- Splitting of labels: some labels need to be split, otherwise they cannot be mapped to other datasets. This applies to conflicts between MVD and Cityscapes labels: *curb* can be *sidewalk* or *terrain*, *bike-lane* and *manhole* can be *sidewalk* or *road*, *rail-track* can be *rail-track* or *road*. Figure 4 shows examples for each category that needs to be split.
- Extension: We introduce two new labels not present in any of the four datasets: *pickup* and *van*. This is done to reduce label confusion as both types appear in several existing classes (see Section 3.3).

All are conceptually visualized in Figure 3 for clarification. This process results in a unified label policy with 80 distinct categories[2].

[2]See supplemental material for a table with all labels and a color legend

| | blur | coverage | distortion | hood | occlusion | overexp. | particles | screen | underexp. | variations |
|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of WD2 frames containing visual hazards (Section 3.1) | | | | | | | | | | |
| low | 43.4% | 16.0% | 9.4% | 16.3% | 34.0% | 6.8% | 4.4% | 33.3% | 5.7% | 5.2% |
| high | 6.0% | 10.6% | 0.1% | 18.9% | 41.0% | 8.2% | 1.9% | 4.1% | 6.7% | 0.5% |
| Impact on PQ / p-value (Section 6.1) | | | | | | | | | | |
| mvd100 | -22.6% | -46.6% | 0.0% | -8.8% | -3.3% | -15.7% | -30.0% | -28.7% | -28.4% | -12.3% |
| | **0.0028** | **0.0002** | 0.0967 | 0.0694 | **0.0202** | **0.0060** | **0.0007** | **0.0015** | **0.0003** | 0.1502 |
| mix150 | -15.5% | -21.0% | 0.0% | -6.3% | -2.6% | -6.7% | -14.8% | -26.3% | -11.0% | -6.1% |
| | 0.0588 | **0.0008** | 0.0914 | **0.0191** | 0.1165 | 0.0595 | 0.0595 | **0.0028** | **0.0057** | 0.1115 |
| Hazard Classifier Performance (Section 6.2) | | | | | | | | | | |
| accuracy | 53.0% | 79.5% | 73.5% | 93.1% | 57.2% | 91.4% | 80.0% | 75.1% | 78.5% | 94.4% |
| macro f1 | 44.2% | 61.2% | 39.1% | 90.4% | 57.2% | 69.2% | 48.0% | 65.5% | 57.7% | 39.1% |

Table 1. Statistics and results relating to visual hazards in the Wilddash2 dataset. Top: Percentage of Wilddash2 frames (public and benchmark) containing specific visual hazards for *low* and *high* severity levels, rest *none*. Middle: Impact of hazards on the average PQ metric of the panoptic segmentation evaluation on the private WD2 benchmark set using the $WD2_{eval}$ label policy. Bold p-values are below the 5% confidence interval and are statistically relevant. Bottom: Accuracy and macro f1-score for the ten prototype hazard classifier.

On the public leaderboard of our dataset benchmark, we use $WD2_{eval}$, a shortened version of our unified label policy. $WD2_{eval}$ consists of 26 classes: the original 19 Cityscapes evaluation labels, the vehicle classes *ego-vehicle*, *pickup*, *van* as well as *billboard*, *streetlight* and *road-marking*. Only vehicle and person classes are considered as instance classes. Negative test cases also evaluate *unlabeled* areas (see Sec. 4.2) This close alignment with the Cityscapes benchmark label policies was chosen to lower the entry barrier for participating users.

### 3.3. Relabeling

The vehicle classes *pickup* and *van* are not found in any of the four datasets. To extend the MVD, Cityscapes, and IDD dataset to our label policy, we manually relabeled their vehicle instances. In addition, the label *autorickshaw* (inspired by the IDD dataset) was also included. Table 2 shows the distribution and source categories for these vehicle classes. The confusion of both vehicle types in category *car* and *truck* was the main motivation to extend the WD2 policy by these new labels.

| Source | van | pickup | autoricks. |
|---|---|---|---|
| MVD car | 4202 (2.8%) | 2654 (1.8%) | 0 |
| MVD other-veh. | 0 | 0 | 128 (8.2%) |
| MVD truck | 43 (0.5%) | 33 (0.4%) | 0 |
| Cityscapes car | 907 (0.6%) | 12 (0.01%) | 0 |
| IDD car | 419 (1.4%) | 10 (0.1%) | - |
| IDD truck | 0 | 18 (0.2%) | - |

Table 2. Addition of *van*, *pickup* and *autorickshaw* class labels. Number of instances and % of source class. Note: Cityscapes and MVD label policies state that pickups should be labelled as *truck*.

### 3.4. Limitations

The new Wilddash2 dataset is specifically designed to cover many visual hazards, but there are some limitations:

- The public sources did not contain frames with strong distortion. Wilddash added a few frames with artificial lens distortion to potentially confuse neural networks. We decided against this approach to preserve the real-world aspect of WD2.
- In many still-images of rain there are either no particles visible or the rain covers the windscreen leading to fewer frames in the *particles* hazard category.
- Out-of-distribution examples for vehicles and people rarely occur. Thus the low number of frames containing the *variations* hazard.

During the development of Wilddash2, the 2.0 update of MVD [21] was introduced. It offers more detailed semantic annotations with added categories and depth ordering cues. However, no new frames were added and no new category addresses any of the label issues presented in this Section. Thus, all information in this work refers to MVD v1.2 but is fully applicable to v2.0 as well.

MSeg [16] scheme targets a similar dataset unification strategy (including non-driving datasets like COCO) without introducing a new dataset themselves. Their policy only includes the reassignment of object labels. This misses cases where outlines of labels need splitting.

Many algorithms use depth data to improve scene understanding performance. However, our method of sourcing frames from public video data does not allow the computa-

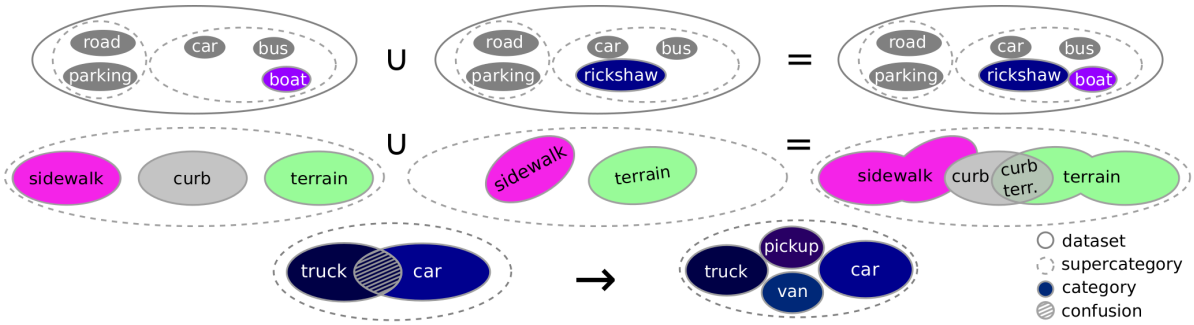


Figure 3. Conceptual depiction of label unification: (top) Organization and combination of disjunct categories and supercategories of two datasets. (center) merging and splitting of sets in case of label-policy-clashes of two datasets (see Figure 4). (bottom) cleaning up mixed categories by the introduction of new label categories.

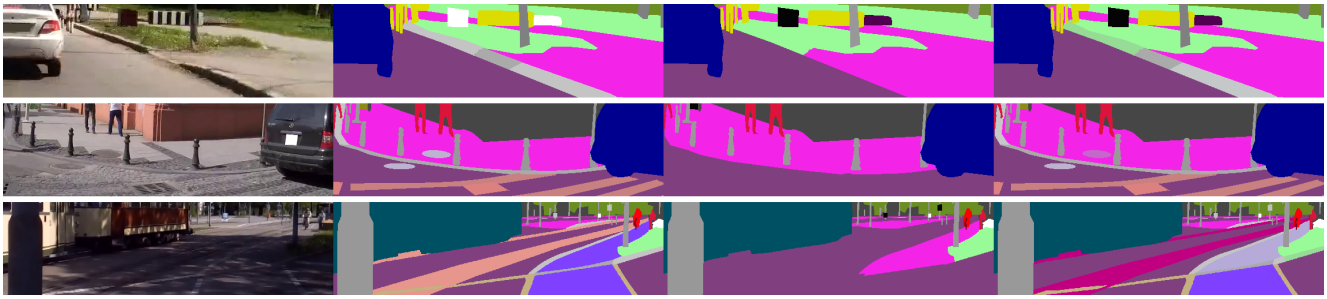

Figure 4. Example frames from WD2 visualizing the need for additional splitting of some labels. Left to right: crop from RGB image, GT using MVD classes, GT using Cityscapes classes, GT using WD2 classes. From top to bottom: ru0009_10000 (*curb* vs. *curb-terrain*), ga0004_10000 (*manhole* vs. *manhole-sidewalk*), de0056_10000 (*bike-lane* vs. *bike-lane-sidewalk* as well as *rail-track* vs. *tram-track*)

tion and release of reliable depth data. This would require a dedicated measurement vehicle, which is contrary to our goal of geographic diversity.

## 4. Evaluation of Panoptic Segmentation

We base our benchmark on the Wilddash public leaderboard which focuses on hard cases and provides more insights using diverse metrics.

Panoptic segmentation [15] describes the combination of instance and semantic segmentation into a single segmentation task. The scene is split into *thing* and *stuff* segments, where *stuff* describes amorphous regions of similar texture (e.g. *road*, *building*) and *thing* describes countable objects (e.g. *person* or *car*). Wilddash2 uses COCO panoptic format [2] for submissions. Panoptic segmentation is evaluated using the *panoptic quality* (PQ) metric defined as follow:

$$PQ = \underbrace{\frac{\sum_{(p,g)\epsilon TP} IoU(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}. \quad (1)$$

Let $g$ be a ground truth segment and $p$ a prediction segment of the same class, $IoU(p,g)$ is the *intersection over union* of the segments $p$ (prediction) and $g$ (GT). A pair of segments $(p,g)$ counts as *true positive* (TP) if the $IoU(p,g)$ is larger than $0.5$. This way, a ground truth segment can only match with at most one prediction segment. The *segmentation quality* (SQ) is the mean $IoU$ of all TP, the *recognition quality* (RQ) penalizes segments without matches, e.g. *false positives* (FP) and *false negatives* (FN).

We apply the concept of hazard-aware testing directly to panoptic segmentation: all metrics are computed separately for the frames from each subset of visual hazards. Impacts per hazard are derived using the method of Zendel *et al*. [39] by comparing results from subsets of different severity levels. Legacy support for both semantic segmentation and instance segmentation is provided: our public toolkit allows the mapping of WD2 into segmentation or instance masks and additional public leaderboards for both tasks help researchers in their respective fields.

### 4.1. Supercategory Scores

Like most panoptic labeling policies, Wilddash2 defines a semantic label on two hierarchical levels:

- an exact identifier that describes the label's specific type (*e.g.* car, truck),
- a broader identifier for label groups (*e.g.* vehicle).

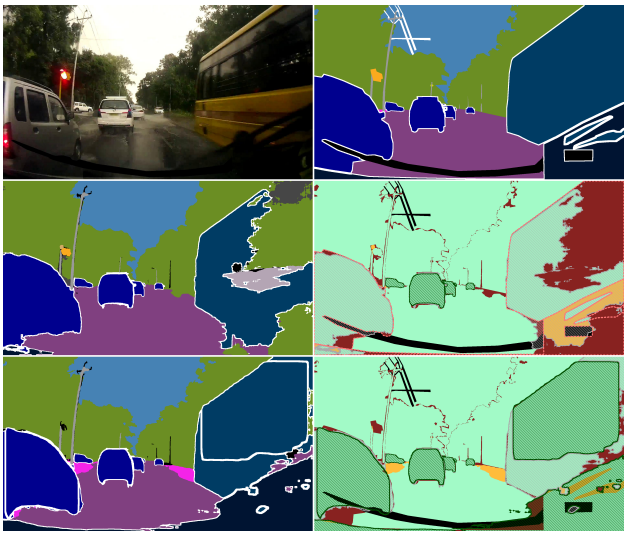

Figure 5. Visualization method for panoptic segmentation results. Top: WD2 scene in0090 RGB image and GT; Middle: result of the MVD-trained model (*mvd100*) and proposed difference image (see Section 4.3); Bottom: result of mixed MVD&WD2 model (*mix150*) and difference image (in0090 was part of the random validation split).

Cityscapes uses the terms *class* and *category*, whereas COCO uses *category* and *supercategory*. To avoid confusion with the term *category*, this paper uses the the terms *category* and *supercategory* for the different hierarchical levels of a semantic label, see Figure 3.

Misclassification of a segment has a negative impact on a model's score. Especially classes that are underrepresented in a model's training set or are annotated differently (e.g. *car* instead of *truck*) are prone to this misclassification. This can skew panoptic evaluation: instances with perfect outlines but wrong category score no points. However, often the wrongly predicted class label and the ground truth share the same supercategory. Wilddash2 extends the evaluation strategy of panoptic segmentation by computing each score (PQ, RQ, SQ) also per supercategory.

From an application perspective, correct supercategory assignments are often more important than overall category correctness. The new supercategory metrics allow additional differentiation between algorithms at a coarser level. In contrast to more complex metrics like $PQ_{Part}$ [5], this is achieved without requiring data relabeling or retraining.

## 4.2. Negative Testing

The Wilddash2 benchmark introduces negative testing to panoptic segmentation. The goal is to evaluate the robustness of a system operated outside of its specifications. Examples from WD2 for such frames include drone scenes, abstract paintings of driving scenes, large-scale image errors, and non-driving scenes (e.g. an indoor volleyball match). Under such circumstances, the desired behavior of a robust system is to mark truly unknown regions as invalid. However, some parts of the image might still contain segments describable by the label policy and systems may be able to produce valid segmentation. The Wilddash2 benchmark rewards the prediction for negative test frames in two ways:

- Reward matching instances: A *best-effort* based on the label policy is defined also for negative test cases. A segment $p$ is detected correctly if the $IoU(g, p)$ with a ground truth segment $q$ of the same *thing* class is larger than 0.5. Correct segments are kept, other segments that overlap with $g$ are set to invalid.
- Reward segments that are flagged as invalid: segment pixels are set to the *best-effort* ground truth, thus improving the overall score of the image.

This combined approach rewards both: systems that create meaningful results for out-of-distribution frames and systems which are aware of their result quality. Existing work on open-set problems (see [12], [23]) focuses on handling gaps in data while our negative testing evaluates systems by investigating their behavior in specific out-of-distribution situations.

Solutions that always "hallucinate" data (*i.e.* never report areas as *unlabeled*) normally have an advantage over more cautious ones: regular metrics potentially only increase by guessing a label, since admitting defeat always lowers the score. Real-world applications are dependent on reliable systems which can estimate the quality of their predictions. Wilddash2 negative testing provides an incentive to encourage improvements in this area.

## 4.3. Visualization

Panoptic segmentation combines semantic per-pixel labels and instancing into a single task. Quantifiable metrics support direct rankings and give a good impression of algorithm performance. Images representing label results can provide a more detailed insight into the workings of a specific solution.

The pure label results themselves can be visualized using standard procedures: false-color mappings represent the labels (*e.g.* light blue for pixels labeled as *sky*) and white outlines encircle individual instances.

Images highlighting the differences between ground truth and predictions help visual inspection of label results. We introduce a novel method to create these "difference images" that illustrates both: the segmentation quality and the instancing quality.

Figure 5 shows visualizations of algorithm results using this method. Segmentation quality is illustrated for pixels with a correct class in mint green, pixels with the false class but correct supercategory in yellow, and pixels with false

| | *MVD* Validation | | | | | *WD2* Benchmark | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PQ$ | $SQ$ | $RQ$ | $PQ_{van}$ | $PQ_{pickup}$ | $PQ$ | $SQ$ | $RQ$ | $PQ_{van}$ | $PQ_{pickup}$ | $PQ_{neg}$ | $PQ_{cat}$ |
| mvd100 | **35.1%** | **74.2%** | **43.9%** | **26.6%** | **29.9%** | 37.6% | 75.6% | 48.3% | 34.0% | 38.1% | 17.1% | 57.7% |
| mix150 | 34.1% | 73.5% | 42.8% | 24.7% | 29.7% | **42.2%** | **77.5%** | **53.2%** | **38.9%** | **49.2%** | **21.1%** | **64.7%** |

Table 3. Performance of the *mvd100* model only trained on MVD for 100 epochs versus *mix150* which is additionally fine-tuned for 50 epochs on WD2. Both evaluated on the original MVD validation set and the hidden WD2 benchmark set. Bold entries mark higher scores.

class and false supercategory in dark red. Areas excluded from comparison receive a black color. The quality of instancing is drawn on top using outlines and hatching. Instances that match a ground truth instance (*i.e.* $IoU(p, q) > 0.5$) are framed and overlaid with a dark green hatched pattern. Wrongly predicted instances (*i.e.* false positives) are framed and overlaid with a grey pattern. Ground truth instances that have no prediction match (*i.e.* false negatives) are framed in a dashed red line and no hatching.

## 5. Statistical Significance

The hazard-aware evaluation method compares performance metrics between subsets of identified hazards, e.g. the performance of an algorithm evaluated at frames marked as having a high severity of occlusions versus frames without occlusions (of instance labels). The quality of such subset comparison can be estimated using a statistical significance test. Such tests work in an inverse fashion: a null hypothesis states that there is no significant difference in subsets and the test should reject this hypothesis in cases where a clear distinction can be made. In our case, the null hypothesis $H_0$ tests that the performance metric is independent of the subset groupings. The significance tests shall reject this $H_0$ hypothesis with a high significance, thus showing that the identified hazard subset is indeed creating a more challenging subset of frames. Demšar [6] offers a good overview of possible statistical significance tests. Initially, no assumption of an underlying distribution of performance metrics can be made. The number of influences on algorithm performance that are present in test frames and how they interact is too complex to estimate. Thus, we chose the non-parametric Mann–Whitney U test [20] to evaluate the significance of hazard subset impacts due to three properties: (1) it does not make assumptions about the underlying distributions (e.g. Gaussian), (2) it does not rely on a direct pairing between individual values, and (3) also works if the subsets have different sample sizes. The test between two subsets for a given metric results in a p-value which is the probability of samples being drawn from the same distribution. A low p-value represents a situation where samples differ strongly and thus the null hypothesis $H_0$ can be rejected. We use a two-sided confidence interval of 5%, i.e. all p-values $< 0.05$ signify that the subsets are substantially different and calculated performance impacts can be trusted.

The results in the middle section of Table 1 include the p-values for each of the visual hazard subsets. The impact of subsets *negative*, *particles*, *occlusion*, *blur*, *screen*, *underexp*, *coverage*, and *overexp* show strong significance. While some hazard evaluations show not enough significance for average metrics, they contain some categories with high significance (e.g. category *ego-vehicle* for subset *hood* or *car* for occlusion). The impacts of *distortion* and *variations* could not be shown with enough significance.

## 6. Experiments

### 6.1. Panoptic Segmentation

The baseline model for panoptic segmentation uses the *Seamless Scene Segmentation* model by Porzi *et al*. [28]. The model *mvd100* is trained using the official BSD-3 codebase [33] on the *Mapillary Vistas* dataset [24] (including relabeled *van* and *pickup* instances) for 100 epochs after which the PQ metric no longer improves on the validation set. The second model *mix150* fine-tunes [3] *mvd100* for additional 50 epochs using a mixture of 3618 randomly selected public Wilddash2 frames (85% of public Wilddash2 frames) and a random subset of 3618 MVD training frames. The remaining 638 public Wilddash2 frames are used as *WD2* validation frames.

Table 3 shows results for both models evaluated on the original MVD validation set and the public Wilddash2 benchmark set (776 frames including 144 negative test cases, GT not public). We show the overall panoptic metrics and individual $PQ$ scores for the newly introduced vehicle classes *pickup* and *van* as well as $PQ$ scores for negative testing and supercategory method as introduced in Section 4. In general, *mix150* is more robust in presence of visual hazards. This comes at the cost of small performance losses for the average MVD frame. The performance reduction for WD2 evaluation of *mvd100* showcases the increased difficulty of WD2.

Table 1 shows the calculated impacts of visual hazards and statistical significance values for each impact (see Section 5). All visual hazards except "distortion" and "varia-

[3] *mvd100* & *mix150* both use MVD labels, see Supplemental for experiments with WD2, Cityscapes, and IDD

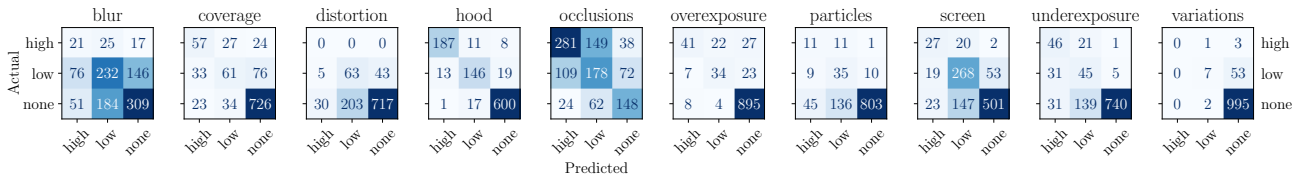


Figure 6. Confusion matrices for each prototype hazard classifiers.

tions" show clearly significant impacts on performance for *mvd100*. As expected, *mix150* suffers a lower performance loss than *mvd100*, proving it to be generally more robust. The confidence for the significance of impact measurements also decreases (higher p-values) for *mix150* signifying a stronger generalization even on hard test cases.

Figure 5 visualizes the output quality of both models for the same frame (used for validation during fine-tuning, i.e. not a training frame).

### 6.2. Visual Hazard Classifiers

The identification of relevant Wilddash2 frames containing visual hazards requires considerable manual effort. Automated hazard classifiers can significantly reduce this work by pre-filtering existing data. Classifiers can potentially also improve the safety of autonomous driving by providing confidence measures for camera-based sensors. First prototypes using the per-image visual hazard meta-labels for each WD2 frame are trained using the *fastai* [10] PyTorch framework. Default augmentations are used to create individual multi-class classifiers per visual hazard based on pre-trained ResNet50 networks. The input resolution of 768x432 and a batch size of 64 are chosen to allow the fast classification of large numbers of video frames. Focal Loss [18] is used to counteract the imbalance of visual hazards subsets and the WD2 full dataset (both public and benchmarking frames) is used to maximize the number of hazards frames. The frames are randomly split into 80% training frames and 20% validation frames The bottom of Table 1 summarizes the classifier performance and Figure 6 shows the respective confusion matrices for all validation frames. The relative low performance of the classifiers *distortion*, *particles*, or *variations* can be accounted to the relative low number of critical cases.

The 5000 frames of WD2 provide sufficient statistical power to identify performance problems for panoptic segmentation but are insufficient to reliably identify visual hazards for arbitrary driving frames. The resulting prototype classifiers successfully perform initial pre-labeling, especially when taking the confidence of the predicted class into account. This reduces the effort for identifying interesting frames by a factor of approx. 10 for the hazards *coverage*, *hood*, *occlusion*, *overexposure*, *screen*, and *underexposure*.

## 7. Conclusion

Panoptic segmentation combines semantic information and individual instancing delivering useful representations for autonomous driving. This work presents the new dataset Wilddash2 which combines the best aspects of four public semantic scene understanding datasets: MVD v1.2, Cityscapes, IDD, and Wilddash. The focus on diverse and difficult scenes complements existing work and with 5000 frames also delivers enough substance for own experiments. Our new data policy with 80 labels is the first to combine the label space of all four datasets and allows precise mapping of WD2 into other domains. Additionally, we identified two new vehicle categories which reduce confusion among instance labels and relabeled all vehicles of MVD, IDD, and Cityscapes. Tools and meta-data for this relabeling are supplied freely under the CC BY-NC-SA 4.0 license thus allowing the inclusion of the new labels in existing frameworks.[4]

We further introduce the concept of hazard-aware testing and negative test cases for panoptic segmentation and provide statistical significance with each performance impact evaluation. This allows for better comparisons and to pinpoint the most pressing issues per algorithm. A new method for visualizing the comparison of panoptic segmentation results helps to quickly understand algorithm characteristics.

Our new public benchmark server with leaderboards allows unbiased comparisons of panoptic segmentation solutions and offers legacy support to evaluate semantic segmentation and instance segmentation as well. The experimental section presents two baseline models showing clear benefits of adding WD2 to your training: increased performance and robustness in visually challenging situations. First prototypes for visual hazard classifiers are presented allowing an automated pre-selection of frames during dataset design. The Wilddash2 dataset and the benchmarking service are available for free to researchers at https://wilddash.cc under CC BY-NC 4.0 license.[5]

[4]This research has received funding from Mobility of the Future; a research, technology, and innovation funding program of the Austrian Ministry of Climate Action

[5]The software for remapping and visualizing panoptic data is released freely under GNU LGPL v2.1 license at https://github.com/ozendelait/wilddash_scripts.

## References


[1] Hermann Blum, Paul Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 2021. 2

[2] COCO - common objects in context. https://cocodataset.org/#format-data. Accessed: 2021-11-01. 5

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 3

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, 2015. 2

[5] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *CVPR*, pages 5485–5494, 2021. 6

[6] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7:1–30, 2006. 7

[7] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi autonomous driving dataset, 2020. 2

[8] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 2

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 1

[10] Jeremy Howard et al. Fastai. https://github.com/fastai/fastai, 2021. Accessed: 2021-10-01. 8

[11] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *CVPRW*, pages 954–960, 2018. 2

[12] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, pages 1175–1184, 2021. 6

[13] ISO Central Secretary. Road vehicles — Safety of the intended functionality. Standard ISO/PAS 21448:2019, International Organization for Standardization, 2019. 2

[14] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. *arXiv preprint arXiv:2103.00868*, 2021. 2

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9404–9413, 2019. 1, 5

[16] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[17] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint arXiv:2109.13410*, 2021. 2

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 8

[19] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 2

[20] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947. 7

[21] Mapillary Research. Mapillary vistas dataset 2.0. https://www.mapillary.com/dataset/vistas. Accessed: 2021-11-12. 4

[22] Kai A. Metzger, Peter Mortimer, and Hans-Joachim Wuensche. A fine-grained dataset and its efficient semantic segmentation for unstructured driving scenarios. In *International Conference on Pattern Recognition (ICPR2020)*, 2021-01. 2

[23] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, pages 3243–3249, 2018. 6

[24] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017. 3, 7

[25] NVIDIA. ClearSightNet. https://news.developer.nvidia.com/drive-labs-helping-cameras-see-clearly-with-ai/. Accessed: 2020-03-02. 2

[26] Andreas Pfeuffer, Markus Schön, Carsten Ditzel, and Klaus Dietmayer. The ADUULM-Dataset - a semantic segmentation dataset for sensor fusion. In *31th British Machine Vision Conference 2020, BMVC 2020, Manchester, UK, September 7-10, 2020*. BMVA Press, 2020. 2

[27] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: Detecting small road hazards for self-driving vehicles. In *IEEE International Conference on Intelligent Robots and Systems*, 2016. 2

[28] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019-06. 7

[29] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2213–2222, 2017. 2

[30] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018. 2

[31] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021-10. 2

[33] Mapillary/Seamseg: Seamless scene segmentation. https://github.com/mapillary/seamseg. Accessed: 2021-11-01. 7

[34] Frederick Tung, Jianhui Chen, Lili Meng, and James J. Little. The raincouver scene parsing benchmark for self-driving in adverse weather and at night. *IEEE Robotics and Automation Letters*, 2017. 2

[35] USGS.gov; Science for a changing world. https://usgs.gov, 2021. Map services and data available from U.S. Geological Survey, National Geospatial Program. Accessed: 2021-10-01. 2

[36] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019. 3

[37] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Pérez. WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9307–9317, 2019. 2

[38] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[39] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *European Conference on Computer Vision (ECCV)*, pages 402–416, 2018. 3, 5

[40] Oliver Zendel, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. CV-HAZOP: Introducing test data validation for computer vision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 1