TU WIEN Informatics

Test Data for Dependable Computer Vision Applications

Cumulative Dissertation

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Dipl.-Ing. Oliver Zendel
Matrikelnummer 0125096

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Dipl.-Ing. Dr.techn. Dr.h.c. Werner Purgathofer
Zweitbetreuung: Dr.techn. Dipl.-Ing. Wolfgang Herzner

Diese Dissertation haben begutachtet:

Anonymized	Anonymized

Wien, 20. April 2023

Oliver Zendel

TU WIEN Informatics

Test Data for Dependable Computer Vision Applications

Cumulative Dissertation

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Oliver Zendel
Registration Number 0125096

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Dipl.-Ing. Dr.techn. Dr.h.c. Werner Purgathofer
Second advisor: Dr.techn. Dipl.-Ing. Wolfgang Herzner

The dissertation has been reviewed by:

Anonymized	Anonymized

Vienna, 20th April, 2023

Oliver Zendel

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Oliver Zendel

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 20. April 2023

Oliver Zendel

Danksagung

Mein Dank gilt meinen Betreuern, meinen Eltern, Familie und meinen Freunden die mich während des Studiums begleitet haben. Ich möchte mich bei allen meinen Kollegen von AIT und HCI für deren Unterstützung und Inspiration durch die zahlreichen Gespräche bedanken.

Acknowledgements

My thanks go to my supervisors, my parents, my family, and my friends for supporting me during my studies. I'd like to thank my colleagues at AIT and HCI for their support and inspiration through the numerous talks and discussions.

Kurzfassung

Das Ziel von Validierung eines bildverarbeitenden Systems ist Zuverlässigkeit und Robustheit in zahlreichen Situationen zu testen. Hierfür werden eigene Testdatensätze verwendet, welche möglichst alle schwierigen Aspekte enthalten, die zum Testen der Robustheit nötig sind. Die Erstellung dieser Datensätze ist aufwendig und schwierig da eine grundlegende Fragestellung offen bleibt: welche Aspekte müssen in den Daten vorhanden sein um Robustheit testen zu können? Diese Doktorarbeit präsentiert einen Lösungsansatz für die Planung neuer Datensätze, das Vergleichen der Qualität von existierenden Datensätzen und die Bestimmung von Lücken in den Daten um diese mit zusätzlichen Testfällen zu stopfen. Der CV-HAZOP Prüfkatalog ist das Resultat einer Risikoanalyse und liefert über 1000 Einträge welche potenziell relevante Aspekte ("visuelle Gefahrenquellen") für Bild-Testdaten identifiziert. Ein Vergleich der Leistung mehrerer Stereo Vision Algorithmen zwischen schwierigen und leichten Bereichen (entsprechend des Prüfkatalogs) zeigen statistisch signifikante Leistungsverluste. Dies bestätigt den Wert des CV-HAZOP Ansatzes um schwierige Situationen zu beschreiben. Bestehende Stereo Vision Datensätze werden analysiert um die Abdeckung und Verteilung von enthaltenen visuellen Gefahrenquellen zu bestimmen. Die Analyse zeigt: bestehende Datensätze konzentrieren sich auf Standard-Situationen und beinhalten nur wenig anspruchsvolle Testfälle. Zur Erstellung eines neuen Datensatzes für semantisches Verständnis von Straßenfahrszenen wird der CV-HAZOP Ansatz angewandt. Der resultierende Testdatensatz Wilddash beinhaltet Fahrszenen aus aller Welt und ein dazugehöriges öffentliches Webservice mit Ranglisten wird erstellt. Dieser Benchmark Service erlaubt das Vergleichen der Robustheit von Algorithmen durch zusätzliches Evaluieren basierend auf Gefahrenquellen sowie Negativtests. Abschließend wird Version 2 von Wilddash präsentiert welches zusätzlich panoptische Segmentierung unterstützt und eine wesentlich höhere Testanzahl beinhaltet. Durch die Erstellung und Anwendung einer neuen vereinigten Kategorisierungs-Regelung ist Wilddash 2 vollständig kompatibel zu drei etablierten Segmentierungsdatensätzen. Eine automatische Erkennung von visuellen Gefahrenquellen mittels Klassifikatoren erlaubt die automatische Vorauswahl von Einzelbildern aus Rohdaten, um die Erstellung von anspruchsvollen Testdatensätzen zu unterstützen.

Abstract

The goal of validation of computer vision (CV) systems is to test the reliability and robustness in various situations. This is done using dedicated test datasets which need to reflect all difficult aspects which the system will potentially face during operation. The creation of such datasets is expensive and difficult while a major challenge remains open: which aspects are actually necessary to test the robustness? This work presents a solution to plan the creation of datasets, compare the quality of existing ones, and pinpoint gaps in the data so that they may be filled using additional test cases. The CV-HAZOP checklist is the result of a risk analysis supplying over 1000 entries which indicate potentially relevant aspects (called "visual hazards") for image test data. Performances of multiple stereo vision algorithms are compared between areas identified as difficult by the checklist vs. regular areas. The statistically significant drop in performance proves the value of this approach in describing challenging aspects. Existing stereo vision datasets are analysed to quantify the coverage of difficult and challenging aspects. This analysis shows: existing datasets focus on standard situations and include only a small amount of visual hazards. The visual hazard approach is then applied to the creation of a new dataset for semantic road scene understanding: Wilddash. A dedicated public benchmark webservice is created which allows the comparison of segmentation algorithm for robustness based on hazard-aware testing and negative testing. Finally, the concept is extended to panoptic segmentation and scaled to match regular state-of-the-art training datasets by creating Wilddash 2. The creation and application of a new unified label policy allows full compatibility of Wilddash 2 with three existing well-known segmentation datasets. The classifier-based detection of visual hazards allows automatic pre-selection of frames to speed up the process of creating challenging large-scale datasets.

Contents

CHAPTER 1

Introduction

This cumulative dissertation contains four papers which present a continuous effort to improve the testing of computer vision systems by creating better test data. The layout is as follows: The motivation Section 1.1 summaries the challenge being tackled by this work. Section 1.1 indicates the goals and Section 1.3 adds a high-level overview of the used methodology covered by all four papers. The overview continues with Section 1.4 which gives individual summaries for each paper. It concludes with Section 1.5 which recounts the contributions in a concise manner. Chapter 2 contains all four papers. Section 3 gives a final summary for the whole dissertation. Section 3.1 presents outlook, some connected papers, and next steps for this work. The original supplemental material for all four papers is included at the end at Section 4.

1.1 Motivation

Computer vision (CV) is a field in computer science where high-level tasks are solved using image and video data. Example tasks include camera-based 3D mapping and localisation, semantic understanding of the environment, and detection of obstacles [Ike21]. These capabilities play a major role in the ongoing automation of many processes in industry and people's daily lives [JGB+20]. Many of these tasks are safety-relevant and mistakes could lead to injury or fatalities. Thus, solutions have to be tested rigorously to ensure their safety. Testing and validation of computer vision systems relies heavily on test datasets. The test cases contained in the datasets include input data (e.g. images, videos) and expected system output (*ground truth* (GT)) and are crucial to detect shortcomings in systems [SHJ11]. They have to include all relevant aspects which the system will face during operation to be effective. Otherwise, unchecked circumstances can occur during system operation and can lead to a dangerous system failure. This presents a huge challenge: how does one assess the quality and completeness of test data?

Gaps in test data represent untested circumstances which can lead to potentially hazardous situations if encountered during a system's life cycle. The open world in which autonomous systems operate can never be completely matched in test data, thus no definitive dataset exists nor a mathematical description of one. Finally, recurrences and redundancies create a weighting during validation: aspects which are common in the data will have a larger impact on the evaluation results compared to minorities [MMS+21].

1.2 Goals

Test data acquisition needs to be planned with these aspects in mind and resulting test cases curated to create test datasets which allow the evaluation of computer vision systems for robustness. The first goal is the systematic collection and cataloguing of situations and aspects which can potentially degrade the performance of a CV system (*visual hazards*).

For specific applications, this results in three main goals:

- Identify gaps in existing test dataset which may lead to dangerous oversight during testing.
- Compare test datasets in regard to the coverage of visual hazards
- Efficiently plan the creation of new test data sets be to include visual hazards.

During the start of this dissertation, no collection of visual hazards existed. No viable strategy or mechanism had been published to plan new test datasets or evaluate existing datasets with visual hazard coverage in mind. This work presents tools and results to improve this situation.

1.3 Methodology

The Section 1.4 summarizes each paper's contributions and the total progress towards the indicated goals. This also includes remarks on the chosen methodology for the individual paper. Important concepts and methods are summarized here for easier comprehension, with respective pages for details.

1.3.1 HAZOP (pages 11- 16, 28- 29)

Hazard and operability analysis [Kle83] (HAZOP) is a risk analysis method. This is a process designed to systematically identify potentially dangerous situations and aspects for a specific system. A HAZOP analysis begins by creating an abstraction of the system that should be investigated. This results in a *model* split into multiple subunits called *locations*. Each subunit is described by *parameters* that control the operation of this component. A standard enumeration of short *guide words* refers to modifications or

deviations of parameters from the expected (e.g. More, Faster, Before). An initial list is generated by combining each guide word with each parameter from the model locations leading to a long list of deviations which could lead to a hazard (i.e. a potential for harm). Hazards in this work's context mostly refer to reduction of sensor data quality, leading to loss of functionality. This in turn reduces the safety of the intended functionality (SOTIF) in contrast to classical hazards, which can result in injury or death for humans. Multiple experts and users working with the real-world system are now tasked to attribute *meaning* to the initial list entries and derive *consequences* for the full system. This is done by contemplating the ways this parameter deviation could result in hazards and *examples* are added to clarify the identified hazards. Results from each participant are collected and discussed to improve consistency and remove duplicates. The HAZOP analysis can result in a thorough organized listing which incorporates the experiences of experts and users guided by the structure it provides.

1.3.2 Stereo Vision (pages 26- 35)

Depth information can be estimated from two synchronized cameras by identifying correlating parts between their images [Mat11]. Software calibration methods are applied to calculate undistorted images representing idealized cameras having parallel central optical axis. Intelligent correlation methods allow the identification of matching pixels between the two images which stem from the same point in the scene. Each correlation can be used to triangulate this scene's point relative to the cameras. This information can be turned to a metric distance measurement by knowing the distance between the cameras and each camera's optical characteristics (intrinsics). Dense stereo vision applies this process to the whole image, thus creating dense distance measures which can be turned into a 3D point cloud of the scene.

1.3.3 Statistical Significance (pages 20, 21, 59)

Evaluations can calculate the expected impacts of a certain visual hazard using test data. The statistical significance of this test gives a measure about how reliable these results can be used to deduce general trend/impact for the system under test. In statistics, a significance can be calculated per specified hypothesis (e.g. performance drops due to glare visible in camera image) and is usually connected to the *p-value* probability[JB19]. This uses an inverse logic: the *p-value* denotes the probability of obtaining the extreme observed values while there is no special relation present (*null hypothesis*, e.g. numbers are uncorrelated or drawn from a random distribution). A low *p-value* indicates that the *null hypothesis* is unlikely and thus there is actually an underlying relation at play (e.g. glare impacts system performace). The threshold below which a *p-value* is considered sufficiently low is called the *significance level* (e.g. 5%).

1.3.4 Concretization (pages 29, 4- 5)

CV-HAZOP resulted in a checklist of potential visual hazard using a model of a generic computer vision application without a specific use case in mind. The process of concretization is transforming this generic list into an interpretation with a specific task in mind (e.g. stereo vision, semantic segmentation). A *task definition* is created that specifies the system's intent. Mirroring the original HAZOP process, this step is conducted by experts and users of the specific application. For each entry in the checklist its meaning, consequences, and examples are re-evaluated in the new concrete context. This includes skipping many entries which do not fit the specific task and will update all included hazards to allow a clear understanding of the hazard in the context of the selected application. Finally, the concretization will result in an updated checklist specialized for the new application.

1.3.5 Negative test cases (pages 28, 31, 32)

Software testing uses a large number of individual test cases to validate a system and identify bugs or unwanted behaviour. Normally a test case represents a use case within the system's specifications resulting in a wanted outcome (e.g. when presented with a certain input, a specified output is expected). For a thorough evaluation, these so-called *positive* test cases can be used in conjunction with *negative* test cases[Sem12, Cem03] which can test system behaviour outside the specification and should result in a specific failure state (e.g. when presented with a certain nonsense-input, the system should report a failure). Positive test cases expect results within the specifications. Thus, any hallucinated or improbable result returned for a difficult scenario will only increase performance compared to admitting to a failure. This can lead to dangerous situations when safety-critical systems interpret unknown scenes. Negative tests are a useful tool to evaluate the robustness of a system by also incentivizing systems to report failure.

1.3.6 Semantic Segmentation (pages 1- 5, 9- 13, 4)

An image is semantically analysed during the computer vision task of semantic segmentation. Each pixel is assigned to one of a limited number of predetermined categories which are relevant for typical use cases[GLGL18]. This attribution of categories to pixels is also called labeling. Examples of useful categories for autonomous driving are: *road*, *sidewalk*, *car*, *traffic sign*, and *person*[COR+16a]. Foreground and background labels are necessary to achieve a full understanding of the scene. Both static scenery and dynamic actors should be included. This semantic information is then used by a higher level algorithm to facilitate tasks such as navigation, path planing, task planing, obstacle avoidance, and visual servoing (vision-guided manipulation tasks).

1.3.7 Panoptic Segmentation (pages 0- 7)

The *per-pixel* semantic labels are extended in the panoptic segmentation task by adding *per-instance* (i.e. per individual unit) information for defined labels[KHG+19]. In the

context of autonomous driving, the vehicle and person classes are typically selected to get additional per-instance labels. Other labels, especially those describing static areas like *grass*, stay on segmentation level as per-instance information would not be useful to fulfill any higher-level task. In a row of parked cars, individual separate labels are created for each car instead of just marking the whole area with a single *car* label (as done for semantic segmentation). Training and evaluation of panoptic segmentation has to weight the quality and errors done per segment class versus the errors done during instancing.

1.4 Paper Summaries

Statement of contribution: Oliver Zendel, the author of this thesis, is the main author for all papers in Section 2 having contributed the largest individual share of time, resources, text, and ideas for each paper. Style, reference sections, and page numbers of the original papers have been preserved. An additional consistent page numbering for this dissertation has been placed at the lower edge corner. Page number ranges in this summary refer to the dissertation page numbering. All attached versions are publicly available open access versions. Some published versions differs in page layout, the acknowledgement section, and copyright notices.

1.4.1 How Good Is My Test Data? Introducing Safety Analysis for Computer Vision

Paper [ZMHH17] at Section 2.1 (pages 9 - 24) lays the foundations for this work: the CV-HAZOP risk analysis. The process is described in detail, resulting in a large checklist of potential visual hazards. The analysed model is a generalization of all computer vision tasks. The concrete example of stereo vision (calculation of dense depth maps using triangulation of two camera images) is used to demonstrate specialization, i.e. the derivation of a checklist for a specific task from the generic CV-HAZOP list. The effect of visual hazards on the resulting quality and robustness of results is compared in an experiment. Three standard stereo vision test datasets are annotated according to the specialized checklist, with bounding boxes marking areas identified by the risk analysis as potentially difficult. The performance comparison of multiple stereo algorithms shows: areas identified as difficult have increased error rates versus various control patches. This comparison is accompanied by a statistical significance analysis to specify the certainty of each evaluation. The risk analysis is a valid method to identify difficult areas. Historically, this paper is a journal version of the conference paper [ZMHH15], extended by a larger experimental section and better descriptions of the CV-HAZOP approach. An earlier paper [ZHM13] presented a first sketch of the whole approach. The full CV-HAZOP checklist with 1470 entries is published freely online [CVH22].

1.4.2 Analyzing Computer Vision Data - The Good, the Bad and the Ugly.

Paper [ZHM+17] at Section 2.2 (pages 25 - 36, supplemental material at 69 - 108) utilizes CV-HAZOP checklist to compare the completeness and difficulty between five major stereo vision test datasets. The analysis shows a critical weakness in existing datasets: they lack many potentially difficult and relevant aspects which can occur during regular operation. Evaluations with potentially too easy test data prevents the detection of flaws and shortcomings within the algorithms. The CV-HAZOP checklist approach reveals shortcomings in existing test data and its verbose description of each visual hazard allows for planning additional test cases to fill specific gaps. Finally, the paper also introduces the distinction of test cases into three classes: positive, borderline, negative. This distinction is often used in regular software testing, but it is still very underrepresented in the field of computer vision.

1.4.3 WildDash - Creating Hazard-Aware Benchmarks

Section 2.3 with paper [ZHM+18] is at pages 37 - 46. It applies the CV-HAZOP checklist to create a whole new test dataset with visual hazards in mind. The new Wilddash dataset and benchmark service for semantic understanding of road scenes is assembled by gathering/filtering test cases to contain previously identified visual hazards. The mapping between test cases and hazards allows calculating the individual performance drop per hazard for a given system. This can be used to characterize and compare algorithms with more detail and helps to pinpoint weaknesses. A suitable method for evaluating negative test cases is introduced, which provides additional feedback for out-of-scope performance. This again creates crucial feedback for the robustness of systems.

1.4.4 Unifying Panoptic Segmentation for Autonomous Driving

Paper [ZSR+22] of Section 2.4 (pages 53 - 63, supplemental material at 109 - 112) vastely expands Wilddash to Version 2 with over 20 times of WD1 test cases. The large sample size improves model training using only WD2 and a unified label policy with 80 label categories allows the combination of WD2 with the three well-known datasets Cityscapes, Mapillary Vistas, and Indian Driving Dataset. Negative testing is extended to the panoptic segmentation metric, which combines two previously disjoint segmentation tasks into one. Image classifiers are trained to create prototypes for automatic visual hazard detectors, which help reduce the manual effort of finding difficult test cases in video material.

1.5 Scientific Contribution

In summary, this work includes these contributions for computer vision research:

- CV-HAZOP: applying risk analysis on computer vision applications as a whole

- Process for creation and evaluation of datasets with visual hazards in mind
- Specialized risk assessment for stereo vision and semantic segmentation tasks
- Evaluation of existing stereo vision datasets regarding challenging aspects
- Creation of a new dataset for understanding road scenes
- Application of negative testing to segmentation tasks
- Application of hazard-aware testing to segmentation tasks
- Creation of an online platform for automated evaluation of road scene understanding
- New visualization methods for results and comparisons of panoptic segmentation
- Proof-of-concept of automated visual hazard detector

The papers included in this thesis have been published at journals and conference proceedings with the highest visibility and impact factors for computer vision research [imp23a, imp23b]. They have already been well received by the community, with a total of over 225 citations [sch23] (status April 2023). The webservice *wilddash.cc* supplying datasets and benchmark service for free to the scientific community has over 2100 registered users from all over the world. Section 3.1 includes summaries of additional projects which are connected to this work.

CHAPTER 2

Papers

2.1 How Good Is My Test Data? Introducing Safety Analysis for CV

Int J Comput Vis (2017) 125:95–109
DOI 10.1007/s11263-017-1020-z

How Good Is My Test Data? Introducing Safety Analysis for Computer Vision

Oliver Zendel[1] · **Markus Murschitz**[1] · **Martin Humenberger**[1] · **Wolfgang Herzner**[1]

Received: 13 April 2016 / Accepted: 18 May 2017 / Published online: 9 June 2017

Abstract Good test data is crucial for driving new developments in computer vision (CV), but two questions remain unanswered: which situations should be covered by the test data, and how much testing is enough to reach a conclusion? In this paper we propose a new answer to these questions using a standard procedure devised by the safety community to validate complex systems: the hazard and operability analysis (HAZOP). It is designed to systematically identify possible causes of system failure or performance loss. We introduce a generic CV model that creates the basis for the hazard analysis and—for the first time—apply an extensive HAZOP to the CV domain. The result is a publicly available checklist with more than 900 identified individual hazards. This checklist can be utilized to evaluate existing test datasets by quantifying the covered hazards. We evaluate our approach by first analyzing and annotating the popular stereo vision test datasets Middlebury and KITTI. Second, we demonstrate a clearly negative influence of the hazards in the checklist on the performance of six popular stereo matching algorithms. The presented approach is a useful tool to evaluate and improve test datasets and creates a common basis for future dataset designs.

Keywords Test data · Testing · Validation · Safety analysis · Hazard analysis · Stereo vision

Communicated by Rene Vidal, Katsushi Ikeuchi, Josef Sivic, Christoph Schnoerr.

✉ Oliver Zendel
oliver.zendel@ait.ac.at

[1] AIT Austrian Institute of Technology, Donau-City-Strasse 1, 1220 Vienna, Austria

1 Introduction

Many safety-critical systems depend on CV technologies to navigate or manipulate their environment and require a thorough safety assessment due to the evident risk to human lives (Matthias et al. 2010). The most common software safety assessment method is testing on pre-collected datasets. People working in the field of CV often notice that algorithms scoring high in public benchmarks perform rather poor in real world scenarios. It is easy to see why this happens:

1. The limited information present in these finite samples can only be an approximation of the real world. Thus we cannot expect that an algorithm which performs well under these limited conditions will necessarily perform well for the open real-world problem.
2. Testing in CV is usually one-sided: while every new algorithm is evaluated based on benchmark datasets, the datasets themselves rarely have to undergo independent evaluation. This is a serious omission as the quality of the tested application is directly linked to the quality and extent of test data. Sets with lots of gaps and redundancy will match poorly to actual real-world challenges. Tests conducted using weak test data will result in weak conclusions.

This work presents a new way to facilitate a safety assessment process to overcome these problems: a standard method developed by the safety community is applied to the CV domain for the first time. It introduces an independent measure to enumerate the challenges in a dataset for testing the robustness of CV algorithms.

The typical software quality assurance process uses two steps to provide objective evidence that a given system fulfills its requirements: verification and validation

(International Electrotechnical Commission 2010). Verification checks whether or not the specification was implemented correctly (i.e. no bugs) (Department of Defense 1991). Validation addresses the question whether or not the algorithm is appropriate for the intended use, i.e., is robust enough under difficult circumstances. Validation is performed by comparing the algorithm's output against the expected results (*ground truth, GT*) on test datasets. Thus, the intent of validation is to find shortcomings and poor performance by using "difficult" test cases (Schlick et al. 2011). While general methods for verification can be applied to CV algorithms, the validation step is rather specific. A big problem when validating CV algorithms is the enormous set of possible test images. Even for a small 8bit monochrome image of 640×480 pixels, there are already $256^{640\times480} \approx 10^{739811}$ possible images). Even if many of these combinations are either noise or render images which are no valid sensor output, exhaustive testing is still not feasible for CV. An effective way to overcome this problem is to find equivalence classes and to test the system with a representative of each class. Defining equivalence classes for CV is an unsolved problem: how does one describe in mathematical terms all possible images that show for example "a tree" or "not a car"? Thus, mathematical terms do not seem to be reasonable but the equivalence classes for images are still hard to define even if we stick to the semantic level. A systematic organization of elements critical to the CV domain is needed and this work will present our approach to supply this.

All in all, the main challenges for CV validation are:

1. What should be part of the test dataset to ensure that the required level of robustness is achieved?
2. How can redundancies be reduced (to save time and remove bias due to repeated elements)?

Traditional benchmarking tries to characterize performance on fixed datasets to create a ranking of multiple implementations. On the contrary, validation tries to show that the algorithm can reliably solve the task at hand, even under difficult conditions. Although both use application specific datasets, their goals are different and benchmarking sets are not suited for validation.

The main challenge for validation in CV is listing elements and relations which are known to be "difficult" for CV algorithms (comparable to optical illusions for humans). In this paper, the term *visual hazard* will refer to such elements and specific relations (see Fig. 1 for examples).

By creating an exhaustive checklist of these visual hazards we meet the above challenges:

1. Ensure completeness of test datasets by including all relevant hazards from the list.
2. Reduce redundancies by excluding test data that only contains hazards that are already identified.

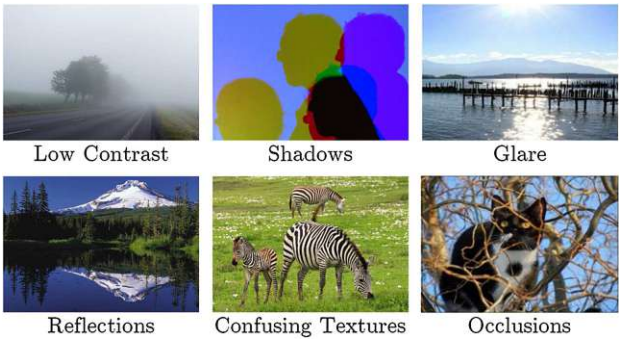

Fig. 1 Examples for potential visual hazards for CV algorithms

Our main contributions presented in this paper are:

- application of the HAZOP risk assessment method to the CV domain (Sect. 3),
- introduction of a generic CV system model useful for risk analysis (Sect. 3.1),
- a publicly available hazard checklist (Sect. 3.7) and a guideline for using this checklist as a tool to measure hazard coverage of test datasets (Sec. 4).

To evaluate our approach, the guideline is applied to three stereo vision test datasets: KITTI, Middlebury 2006 and Middlebury 2014 (see Sect. 5). As a specific example, the impact of identified hazards on the output of multiple stereo vision algorithms is compared in Sect. 6.

2 Related Work

Bowyer and Phillips (1998) analyze the problems related to validating CV systems and propose that the use of sophisticated mathematics goes hand in hand with specific assumptions about the application. If those assumptions are not correct, the actual output in real-world scenarios will deviate from the expected output.

Ponce et al. (2006) analyze existing image classification test datasets and report a strong database bias. Typical poses and orientations as well as lack of clutter create an unbalanced training set for a classifier that should work robustly in real-world applications.

Pinto et al. (2008) demonstrate by a neuronal net, used for object recognition, that the currently used test datasets are significantly biased. Torralba and Efros (2011) successfully train image classifiers to identify the test dataset itself (not its content), thus, showing the strong bias each individual dataset contains.

A very popular CV evaluation platform is dedicated to stereo matching, the Middlebury stereo database. Scharstein and Szeliski (2002) developed an online evaluation platform which provides stereo datasets consisting of the image pair and the corresponding GT data. The datasets show indoor scenes and GT are created with a structured light approach (Scharstein and Szeliski 2003). Recently, an updated and enhanced version was presented which includes more challenging datasets as well as a new evaluation method (Scharstein et al. 2014). To provide a similar evaluation platform for road scenes, the KITTI database was introduced by (Geiger et al. 2012).

A general overview of CV performance evaluation can be found in (Thacker et al. 2008). They summarize and categorize the current techniques for performance validation of algorithms in different subfields of CV. Some examples are shown in the following: Bowyer et al. (2001) present a work for edge detection evaluation based on receiver operator characteristics (ROCs) curves for 11 different edge detectors. Min et al. (2004) describe an automatic evaluation framework for range image segmentation which can be generalized to the broader field of region segmentation algorithms. In Kondermann (2013) the general principles and types of ground truth are summarized. They pointed out, that thorough engineering of requirements is the first step to determine which kind of ground truth is required for a given task. Strecha et al. (2008) present a multi-view stereo evaluation dataset that allows evaluation of pose estimation and multi-view stereo with and without camera calibration. They additionally incorporate GT quality in their LIDAR-based method to enable fair comparisons between benchmark results. Kondermann et al. (2015) discuss the effect of GT quality on evaluation and propose a method to add error bars to disparity GT. Honauer et al. (2015) reveal stereo algorithm-specific strengths and weaknesses through new evaluation metrics addressing depth discontinuities, planar surfaces, and fine geometric structures. All of these are examples of visual hazards.

Current test datasets neither provide clear information about which challenges are covered nor which issues remain uncovered. Our approach can fill both gaps: By assigning a reference-table entry with a unique identifier to each challenging hazard, we create a checklist applicable to any dataset. To the best knowledge of the authors there is no published work considering the vision application as a whole, which identifies risks on such a generic level.

2.1 Robustness

Depending on context, *robustness* can refer to different characteristics of the considered system. In the safety context, robustness is about the correct handling of abnormal situations or input data. For instance, in the basic standard for functional safety (International Electrotechnical Commission 2010), it is defined via the system's behavior in a hazardous situation or hazardous event. This also includes the ability to perform a required function in the presence of implementation faults (internal sources) or cope with faulty and noisy input data (external sources). A method to evaluate robustness against the first type is fault injection (Hampel 1971) (e.g., bit flips in registers or bus) while fuzz testing (Takanen et al. 2008) can be used for assessing the robustness against abnormal input data.

In computer vision, robustness usually refers to coping with distorted or low-quality input. Popular methods are random sample consensus (RANSAC) (Fischler and Bolles 1981), M-Estimators (Huber 1964), or specific noise modeling techniques, which arose from the need to use systems in "real-world applications". In the work described in this paper, we do not exclude these issues, but aim to cover all influences that may cause a degraded or false performance of a CV solution. This in particular includes aspects that can usually be part of observed scenes, such as lacking or highly regular textures, reflections, occlusions, or low contrasts. Figure 1 illustrates some examples.

2.2 Risk Analysis

Risk-oriented analysis methods are a subset of validation and verification methods. All technical risk analysis methods assess one or several risk-related attributes (e.g. safety or reliability) of systems, components or even processes with respect to causes and consequences. Some techniques additionally try to identify existing risk reduction measures and propose additional measures where necessary.

Originally, risk identification techniques have been developed by the chemical industries, but nowadays they are successfully applied to software quality assurance as well (see Fenelon and Hebbron 1994 and Goseva-Popstojanova et al. 2003 for UML models). The most commonly used methods are:

- HAZOP [7], (Kletz 1983)—hazard and operability analysis,
- FME(C)A (Department of Defense 1949)—failure modes, effects, (and criticality) analysis,
- FTA (Vesely et al. 1981; Laprie 1992)—fault tree analysis.

Each risk analysis method defines a systematic process to identify potential risks. The first step in a HAZOP is to identify the essential components of the system to be analyzed. The parameters for each component, which define its behavior, have to be identified. These parameters often describe the input output characteristics of the component. A set of predefined guide words which describe deviations

are applied to the parameters (e.g. “less” or “other than”) and the resulting combinations are interpreted by experts in order to identify possible consequences (potential hazards) and counteractions. While FME(C)A also starts with identifying the systems components and their operating modes, it then identifies the potential failure modes of the individual components. Further steps deal with identifying potential effects of these failures, their probability of occurrence, and risk reduction measures similar to HAZOP. FTA starts with a hazardous “top event” as root of the fault tree. Leaves are added recursively to the bottom events representing Boolean combinations which contain possible causes for their parent event (e.g. “own car hits the front car” if “speed too high” and “braking insufficient”). This refinement is executed until only elementary events are encountered.

3 CV-HAZOP

The identification and collection of CV hazards should follow a systematic manner and the results should be applicable to many CV solutions. The process has to be in line with well-established practices from the risk and safety assessment community to create an accepted tool for validation of CV systems. The most generic method HAZOP (Kletz 1983) is chosen over FME(C)A and FTA because it is feasible for systems for which little initial knowledge is available. In addition, the concept of guide words adds a strong source of inspiration that all other concepts are missing.
The following Sections address the main steps of a HAZOP:

1. Model the system.
2. Partition the model into subcomponents, called locations.
3. Find appropriate parameters for each location which describe its configuration.
4. Define useful guide words.
5. Assign meanings for each guide word/parameter combination and derive consequences from each meaning.
6. Give an example clarifing the entry using for a specific application (e.g. in the context of stereo vision, object tracking, face detection).

3.1 Generic Model

The first step of any HAZOP is deriving a model of the system that should be investigated. In case of this HAZOP, the generic CV algorithm has to be modeled together with the observable world (its application). Marr (1982) proposes a model for vision and image perception from the human perception perspective. Aloimonos and Shulman (1989) extended it by the important concepts of stability and robustness. We propose a novel model which is entirely based on the idea of information flow: The common goal of all CV algorithms is the extraction of information from image data. Therefore “information” is chosen to be the central aspect handled by the system. It should be noted, that “information” is used in the context “Information is data which has been assigned a meaning.” Van der Spek and Spijkervet (1997) rather than in a strict mathematical sense (Shannon and Weaver 1949). In this context, *hazards are all circumstances and relations that cause a loss of information*. Even though hazards ultimately propagate to manifest themselves in the output of the algorithm, an effective way to find a feasible list of hazards is to look at the entire system and attribute the hazard to the location where it first occurred (e.g. unexpected scene configuration or sensor errors). Multiple inputs from different disciplines are used to create the system model:
Information Theory Communication can be abstracted according to information theory (Shannon and Weaver 1949) as information flow from the transmitter at the source—with the addition of noise—to the receiver at the destination.
Sampling Theorem Sampling is a key process in the course of transforming reality into discrete data. Artifacts that can be caused by this process, according to (Nyquist 1928; Shannon 1949), will result in a loss of information.
Rendering Equation The rendering equation (Kajiya 1986) is a formal description of the process of simulating the output of a virtual camera within a virtual environment. The different parts of the standard rendering equation amount to the different influences that arise when projecting a scenery light distribution into a virtual camera.
Control Theory The general system theory (e.g. Von Bertalanffy 1968) and especially cybernetics interpret and model the interactions of systems and the steps of acquiring, processing, as well as reacting to information from the environment.
The entire flow of information is modeled as follows:

1. Since in CV the sensor is a camera, all data within the observed scene available to a CV component can only be provided by the electromagnetic spectrum (simply referred to as *light* in this paper) received by the *observer* (i.e. the sensor/camera) from any point in the scene. Hence, light represents data and, as soon as a meaning is assigned, information.
2. At the same time, any unexpected generation of light and unwanted interaction of light with the scene distorts and reduces this information.
3. The sensing process, i.e. the transformation of received light into digital data, further reduces and distorts the information carried by the received light.
4. Finally, the processing of this data by the CV algorithm also reduces or distorts information (through rounding errors, integration etc.).

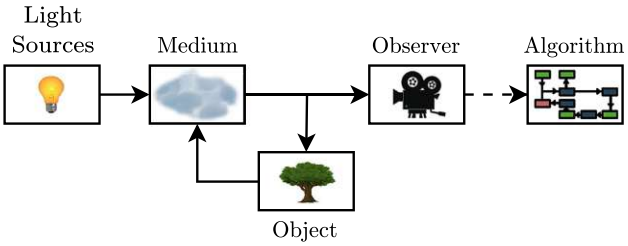

Fig. 2 Information flow within the generic model. Light information is shown as *solid arrows* and digital information as a *dashed arrow*

In essence, two information carriers are distinguished: light outside of the system under test (SUT) and digital data within the SUT. This is visualized in Fig. 2 by two types of arrows: solid arrows for light and a dashed line for digital data. At each transition, this information can potentially be distorted (e.g. by reduction, erasure, transformation, and blending). Benign interactions, e.g. interaction of a pattern specifically projected to create texture for structured light applications, are not entered into the list. We are interested in situations and aspects that can potentially reduce output quality. Nevertheless, the failing of such expected benign interactions (e.g. inference effects of multiple projectors) are a risk and, thus, included in the analysis.

3.2 Locations

The system model is now partitioned into specific locations (i.e. subsystems) of the overall system. Light sources that provide illumination start the process flow (illustrated in Fig. 2). The light traverses through a medium until it either reaches the observer or interacts with objects. This subprocess is recursive and multiple interactions of light with multiple objects are possible. The observer is a combination of optical systems, the sensor, and data pre-processing. Here the light information is converted into digital data as input for a CV algorithm. The CV algorithm processes the data to extract information from it.

Each entity (box in Fig. 2) represents a location for the HAZOP. The recursive loop present in the model results in an additional location called “Objects” for aspects arising from the interactions between multiple objects. The observer is modeled by two components: “Observer—Optomechanics” and “Observer—Electronics”. This reduces complexity for the analysis and allows to focus on the different aspects of the image capturing process.

3.3 Parameters

Each location is characterized by parameters. They refer to physical and operational aspects describing the configuration of the subcomponent. The set of parameters chosen for a single location during the HAZOP should be adequate for its characterization. Table 2 shows the parameters chosen for the location “Medium” as an example. Too few parameters for a location means that it is insufficiently modeled and that the analysis will likely contain gaps. Performing an analysis with too many parameters would require too much effort and create redundancy. A full listing of all parameters is available at the website vitro-testing.com.

3.4 Guide Words

A guide word is a short expression to trigger the imagination of a deviation from the design/process intent. Number and extent of guide words must be selected to ensure a broad view on the topic. Nevertheless, their number is proportional to the time needed for performing the HAZOP, so avoiding redundant guide words is essential. The provided examples in Table 1 show all guide words we used in the analysis. Exemplary meanings for the deviations caused by each guide words are given, but the experts are not limited to these specific interpretations during the risk analysis. The first seven “basic” guide words are standard guide words used in every HAZOP. The remainder are adaptations and additions that provide important aspects specific for CV: spatial and temporal deviations (Table 2).

3.5 Implementation

The actual implementation of the HAZOP is the systematic investigation of each combination of guide words and parameters at every location in the system. It is performed redundantly by multiple contributors. Afterwards, the results are compared and discussed to increase quality and completeness. Each HAZOP contributor assigns at least one meaning to a combination. In addition, for each meaning found the contributors investigate the direct consequences of this deviation on the system. One meaning can result in multiple consequences at different levels. Each entry in the list represents an individual hazard which can lead to actual decreases in the total system’s performance or quality. Combinations that result in meaningful interpretations by any contributor are considered to be “meaningful” entries while combinations without a single interpretation are considered to be “meaningless”.

3.6 Execution

The execution of the CV-HAZOP, including various meetings and discussions by the contributors (with expertise in testing, analysis, and CV), took one year. Each location is covered by at least three of the authors. The additional experts are mentioned in the acknowledgments. The 52 parameters from all seven locations, combined with

Table 1 Guide Words used in the CV-HAZOP

Guide word	Meaning	Example
Basic		
No	No information can be derived	No light at all is reflected by a surface
More	Quantitative increase (of parameter) above expected level	Spectrum has a higher average frequency than expected
Less	Quantitative decrease below expected level	Medium is thinner than expected
As well as	Qualitative increase (additional situational element)	Two lights shine on the same object
Part of	Qualitative decrease (only part of the situational element)	Part of an object is occluded by another object
Reverse	Logical opposite of the design intention occurs	Light source casts a shadow instead of providing light
Other than	Complete substitution—another situation encountered	Light source emits a different light texture
Additional—spatial		
Where else	"Other than" for position/direction related aspects	Light reaches the sensor from an unexpected direction
Spatial periodic	Parameter causes a spatially regular effect	A light source projects a repeating pattern
Spatial aperiodic	Parameter causes a spatially irregular effect	The texture on object shows a stochastic pattern
Close/remote	Effects caused when s.t. is close to/remote of s.t. else	Objects at large distance appear too small
In front of/behind	Effects caused by relative positions to other objects	One object completely occludes another object
Additional—temporal		
Early/Late	Deviation from temporal schedule	Camera iris opens too early
Before/after	A step is affected out of sequence, relative to other events	Flash is triggered after exposure of camera terminated
Faster/slower	A step is not done with the right timing	Object moves faster than expected
Temporal periodic	Parameter causes a temporally regular effect	Light flickers periodically with 50 Hz
Temporal aperiodic	Parameter causes a temporally irregular effect	Intensity of light source has stochastic breakdowns

Table 2 Parameters used in the location Medium

Parameter	Meaning
Transparency	Dimming factor per wavelength and distance unit
Spectrum	Color, i.e. richness of medium with respect to absorption spectrum (isotropic or anisotropic)
Texture	Generated by density fluctuations and at surfaces (e.g. water waves)
Wave properties	Polarization, coherence
Particles	Influences and effects of the particles that make up the medium

the 17 guide words, result in 884 combinations. Each combination can have multiple meanings assigned to it. Finally, 947 unique and meaningful entries have been produced. Table 3 shows an excerpt of entries from the final HAZOP and Fig. 3 shows visualizations for each hazard mentioned. The entries in the list can include multiple meanings for each parameter as well as multiple consequences and hazards per meaning. The whole resulting dataset of the CV-HAZOP is publicly available at www.vitro-testing.com.

Table 3 Excerpt from CV-HAZOP entries (simplified)

HID	Location/parameter	Guide word	Meaning	Consequence	Example
16	Light sources/position	No	Position of light source not known	Light source is not detected	Confusion of algorithm due to missing relation of shadows to the respective light source
271	Medium/particles	Faster	Particles move faster than expected	Motion blur of particles	Blurred particles obfuscate scene; image recognition severely reduced
370	Object/complexity	More	Object is more complex than expected	Object has features or feature combinations which make its correct recognition difficult	Object is not correctly recognized
537	Objects/number	No	Number of objects is not detectable/decidable	Scene with unknown number of objects	Nonexistent objects are reported (false positives)
724	Objects/transparency	More	Objects are more transparent than expected	More objects are visible than usual	Algorithm is confused by clutter in scene
997	Observer/transmittance of optics	Spatial periodic	The optical density of the lenses changes in a spatially periodic manner, e.g., microlens array	The scene appears periodically repeated over the image in shrunk size	Strong confusion of CV alg.
1154	Observer/resolution	More	The sensor resolution is higher than expected	More noise than expected	Increase error rate on scene interpretation due to higher noise rate
1350	Algorithm/models	Part of	Models are incomplete	System uses incomplete model data	Misdetections due to incomplete models

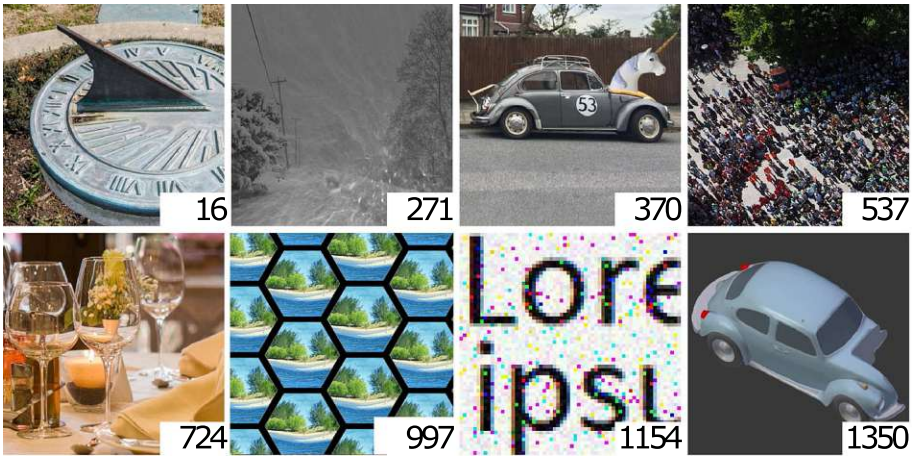

Fig. 3 Visualization of entries from Table 3

3.7 Resulting List

In total, 947 entries are considered meaningful by the experts. A detailed analysis of the meaningful entries achieved for each guide word/parameter combination is shown in Fig. 4. One goal is to maximize the meaningful entries—and the graphic shows reasonably high entries for most of the basic guide words (see Table 1). Lower valued entries in the matrix can be explained as well: The concepts of the spatial aspects "Close" and "Remote" are simply not applicable to the properties of the electronic part of the observer (obs. electronics) and the concept of space in general is not applicable to a number of parameters at various locations. This also holds true for the temporal guide words which do not fit to the optomechanical and medium locations. Nevertheless, even here the usage of guide word/parameter combinations inspire the analysts to find interpretations which would have been hard to find otherwise. Each hazard entry is assigned a unique hazard identifier (*HID*) to facilitate referencing of individual entries of the checklist.

4 Application

The remainder of this paper focuses on the application of the checklist as an evaluation tool for existing test datasets. On the one hand, we show that the CV-HAZOP correctly identifies challenging situations and on the other hand, we provide a guideline for all researches to do their own analysis of test data.

Initially, the evaluators have to clarify the intent and domain of the specific task at hand. This specification creates the conceptual borders that allow the following analysis to filter the hazards. The intent includes a description of the goals, the domain defines the conditions and the environment under which any algorithm performing the task should work robustly. With the intent and domain specified, the evaluators can now check each entry of the CV-HAZOP list to see if that entry applies to the task at hand. Often it is useful to reformulate the generic hazard entry for the specific algorithm to increase readability. In the following a process outline is given:

	Generic							Temporal							Spatial					
	No	More	Less	As well as	Part of	Reverse	Other than	Temp. periodic	Temp. aperiodic	Before (temp.)	After	Faster	Slower	Where else	Spat. periodic	Spat. aperiodic	Close (spat.)	Remote (spat.)	In front of (spat.)	Behind (spat.)
Light Source	1.00	1.00	1.00	1.00	0.63	0.56	0.78	1.00	1.00	0.67	0.73	0.60	0.25	0.67	0.88	0.75	0.63	0.38	0.56	0.56
Medium	1.00	1.00	1.00	1.00	0.60	0.80	1.00	0.80	0.80	0.20	0.20	0.83	0.80	0.80	1.00	1.00	0.80	1.00	0.40	0.40
Object	1.00	1.00	1.00	1.00	0.91	0.91	1.00	1.00	1.00	1.00	1.00	0.90	0.60	0.50	0.91	0.92	0.82	0.73	0.30	0.30
Objects	1.00	1.00	1.00	1.00	1.00	0.82	0.88	1.00	1.00	0.71	0.75	1.00	0.71	0.92	1.00	0.86	1.00	1.00	1.00	1.00
Obs. Optomechanics	1.00	1.00	1.00	0.93	1.00	0.75	1.00	0.69	0.86	0.64	0.69	0.75	0.62	0.92	0.85	0.83	0.75	0.67	0.50	0.58
Obs. Electronics	1.00	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.83	1.00	1.00	1.00	0.80	1.00	1.00	0.00	0.00	0.20	0.20
Algorithm	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.57	0.50	0.71	0.83	0.33	0.33

Fig. 4 Ratio of meaningful combinations for each guide word per location (averaged over all parameters of each location)

1. Check if the preconditions defined by the column *Meaning* and the according *Consequences* apply.
2. Check if the *Example* matches the specific task at hand.
3. Each row represents a unique *Hazard* and has a unique *Hazard ID* (HID). If the *Hazard* is too generic to be feasible, add a new row for the specific task using a matching *Example*.
4. Evaluate if the *Hazard* can be detected (i.e. is visible in the test data).
5. Store the identity of test cases which fulfill relevant *HIDs*. Create a new test case should none of the current test cases fulfill this *Hazard*.

Previous evaluations for comparable tasks can be used as templates to speed up this process and to reduce the effort compared to evaluating the whole generic list. Specialized hazards can be added to the checklist so that they can be used directly in future evaluations.

With the reduced list of possible hazards, the evaluators are able to go through test datasets and mark the occurrence of a hazard. Usually a simple classification per test case is enough. Individual pixel-based annotations can also be used to indicate the location of specific hazards in test images (see Sect. 5). After this process, the missing hazards are known and quantifiable (e.g. 70% of all relevant hazards are tested using this test dataset). This is a measure of completeness which can be used to compare datasets. Even more important: If a hazard cannot be found in the test data, the CV-HAZOP entry states an informal specification for creating a new test case to complement the test dataset. The extensiveness of the checklist allows a thorough and systematic creation of new test datasets without unnecessary clutter.

Each hazard entry in the check list has a unique hazard identifier (HID). This allows to easily reference individual hazards and compare results from different CV implementations. The checklist approach allows for a top-down evaluation of CV (starting from the problem definition down to the pixel level). This is a good complement to regular benchmarks which tend to be focused on the detailed pixel level (bottom-up evaluation).

5 Example

As proof of concept, the authors applied the described process to a specific task. We chose canonical stereo vision:

The intent of the algorithm is the calculation of a dense disparity image (correspondence between the pixels of the image pair) with a fixed epipolar, two camera setup. To further simplify the analysis, we only use greyscale information and assume that the cameras are perfectly synchronous (exposure starts and stops at the same instants), and omit the use of any history information so that many time artifacts can be disregarded. The domains of the algorithm are indoor rooms or outdoor road scenarios. Conditions like snow, fog, and rain are included in the problem definition. This was done to keep the problem definition sufficiently generic to allow room for the analysis.

Note that this evaluation is not designed to compare stereo vision algorithms themselves or to compare the quality of the specific datasets (will be done in future works). However, this paper provides the first step: a clear proof of concept of the CV-HAZOP list as a tool for validation. The simplifications in domain/intent analysis and algorithm evaluation were performed to reduce complexity/workload and should be re-engineered for a specific stereo vision evaluation.

First, six experts in the field of CV (some had experience with the CV-HAZOP list, others were new to the concept) analyzed the initial 947 entries and identified those applying to the stereo vision use case. During this step, 552 entries were deemed to be not applicable and 106 entries were non-determinable (not verifiable by only surveying the existing test data; more background knowledge needed). The remaining 289 entries were deemed to be relevant for stereo vision. See Table 4 and Fig. 5 for examples from the datasets. About 20% of the hazard formulations were further specified to simplify the following annotation work while the rest were already specific enough. The experts analyzed three test datasets commonly used for stereo vision evaluation (see Table 5) individually for each of the identified hazard.

The hazard entries were evenly distributed among evaluators. All evaluators had the task to annotate each assigned hazard at least once in each dataset (if present at all). The step to annotate all occurrences of individual hazards in all images was omitted as the required effort would exceed the resources reasonable for this proof of concept. One representative of each hazard is deemed sufficient for the purpose of this proof-of-concept but certainly requires a larger sample size for a detailed evaluation of a CV algorithm. Ideally, a test dataset should include a systematically increasing influence of each hazard so that the algorithm's point of failure can be evaluated.

The annotation tool was set to randomly choose the access order to reduce annotation bias by removing the influence of image sequence ordering. Table 5 summarizes the results of the evaluation showing the number of images with hazards and the number of uniquely identified hazards. It is not a surprise that KITTI contains the most hazards: it is the largest dataset and is also created in the least controlled environment (outdoor road scenes). It contains many deficiencies in recording quality manifesting as hazards and it includes images with motion blur as well as reflections on the windshield.

Many effects stemming from interactions of multiple light sources, medium effects, and sensor effects are missing in all three test datasets. The majority of hazards present in the data

Table 4 Example for entries used during the example application of the CV-HAZOP list (simplified)

HID	Location/parameter	Guide word	Meaning	Consequence	Example
125	Light source/intensity	More	Light source shines stronger than expected	Too much light in scene	Overexposure of lit objects
481	Object/reflectance	As well as	Obj. has both shiny and dull surface	Diffuse reflection with highlight/glare	Object recognition distorted by glares
445	Object/texture	No	Object has no texture	Object appears uniform	No reliable correspondences can be found
706	Objects/reflectance	Close	Reflecting Obj. is closer to Observer than expected	Reflections are larger than expected	Mirrored scene taken for real
584	Objects/positions	Spatial periodic	Objects are located regularly	Same kind of objects appear in a geometrically regular pattern	Individual objects are confused
1059	Optomechanics/aperture	Where else	Inter-lens reflections project outline of aperture	Ghosting appears in the image	Aperture projection is mis-interpreted as an object
1123	Electronics/exposure	Less	Shorter exposure time than expected	Less light captured by sensor	Details uncorrelated due to underexposure

deal with specific situations that produce overexposure (HIDs 26, 125, 479, 482, 655, 707, 1043, 1120), underexposure (HIDs 21, 128, 651, 1054, 1072, 1123), little texture (HIDs 444, 445, 449) and occlusions (HIDs 608, 626).

6 Evaluation

In this section we evaluate the effect of identified hazards on algorithm output quality. The goal is to show that the entries of the CV-HAZOP are meaningful and that the checklist is a useful tool to evaluate robustness of CV algorithms. A specific hazard can only impact the system if it is visible in the image. Thus, we need to annotate areas in images corresponding to specific hazards to show that the annotated area itself (and, thus, the included hazard) is responsible for the output quality decrease. Initially it was unclear how accurate these areas have to be defined. For this purpose two different types of annotations were evaluated: a manually selected outline and a bounding box calculated from the outline.

We potentially add another bias to our analysis by evaluating only areas that contain annotations. This has two influences: (i) We only look at frames that have annotations while ignoring all other frames in the dataset without any annotations, (ii) We average over small sampling windows that often contain relatively little data due to missing values in the GT.

To quantify these influences we generated another set of control annotations: for each annotation in the dataset we generated a mask with a random position but the same size as the annotated hazard in the respective frame.

At last the overall performance of an algorithm was needed as a base line value. For this the whole image was evaluated. All in all we generated four types of masks from the annotations for our evaluation.

The different masks represent a step-by-step increase of influence of the annotated areas:

- *shape* masks with the annotated outlines as filled polygons,
- *box* masks with boxes of equal size and centroid as each annotated outline,
- *rand* masks with boxes of equal size as the annotated outlines but a randomly placed centroid,
- *all* masks with all pixels except the left border region (to exclude occlusions).

Figure 6 gives an example of the generated masks. Not every image in the test datasets contains annotations. The masks *shape*, *box*, and *rand* are evaluated for the subset of images containing at least one annotation while *all* is evaluated for all images of the datasets.

Fig. 5 Examples for each entry in Table 4. Images are taken from the datasets described in Table 5

Table 5 Stereo vision test datasets used in our evaluation, number of found hazards and percentage of masks covered by GT

Algorithm	Image pairs	Images with hazards	Found hazards	# Annotations	% GT all	% GT rand (avrg.)	% GT box	% GT shape
Middlebury Stereo Evaluation (MB06) (Scharstein and Szeliski 2002)	26	19	34	55	96.0	88.2	94.8	92.5
Middlebury Stereo Eval. "New" (MB14) (Scharstein et al. 2014)	23	17	57	80	96.9	93.0	93.5	91.2
The KITTI Vision Benchmark (KITTI) (Geiger et al. 2012)	194	62	76	101	45.7	42.0	36.8	37.0

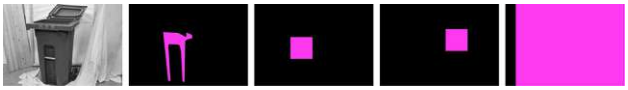

Fig. 6 Example for annotation masks for hazard 'No Texture' (from *left to right*): input image, shape, box, rand, all

The *rand* masks only represent the annotated area's size as well as the subset of annotated frames. A total of 100 random masks are generated for each annotation that share its size but are randomly displaced. Statistics can thus be evaluated over the whole set of random masks which increases the significance. Annotation *box* represents area and position while *shape* represents the full annotation.

The *rand* versus *all* masks verify if the output quality is affected by using smaller image parts for evaluation instead of the whole image as well as a subset of frames, while *box* versus *shape* evaluates the influence of specific shapes of the annotations.

Table 5 lists the resulting number of annotations created for each dataset. Some hazards require the selection of split areas, resulting in multiple annotations. We only use pixels with valid GT information for evaluation. Unfortunately, many of the hazards (e.g. reflections, transparencies, occlusions, very dark materials) also have a negative influence on the laser scanner used for the GT generation in KITTI. The GT data is generally sparse and even more sparse in the annotated areas.

6.1 Performance Evaluation

For evaluation of the stereo vision test dataset we used the following popular stereo vision algorithms: SAD + texture thresholding (TX) & connected component filtering (CCF) (Konolige 1998), SGBM + TX & CCF (Hirschmüller 2008), census-based BM + TX & CCF (Humenberger et al. 2010; Kadiofsky et al. 2012), cost-volume filtering (CVF) & weighted median post processing filtering (WM) (Rhemann et al. 2011), PatchMatch (PM) & WM (Bleyer et al. 2011), and cross-scale cost aggregation using census and segment-trees (SCAA) & WM (Zhang et al. 2014), (Mei et al. 2013). The resulting disparities of each stereo vision algorithm are compared to the GT disparities of the test dataset. The number of wrong pixels (with an error threshold of >2px) is then compared to the number of pixels within the respective mask that had valid ground truth values. Invalids in the result are counted as being above any threshold. We consider each disparity pixel $d_i \in \mathbb{R}^\star$ to either be valid ($\in \mathbb{R}$) or invalid (denoted by the star value "$\star$"). Where $\mathbb{R}^\star = \mathbb{R} \cup \{\star\}$. The same holds for each corresponding ground truth pixel value $g_i \in \mathbb{R}^\star$. We consider every d_i for which $correct(d_i, g_i) = true$ to be true, and $correct : \mathbb{R}^\star \times \mathbb{R}^\star \mapsto$ true, false to be defined by:

$$correct(g_i, d_i) = \begin{cases} true & for\ d_i = \mathbb{R} \\ & \bigwedge\ g_i \neq \star \\ & \bigwedge\ |d_i - g_i| < 2 \\ false & else \end{cases} \quad (1)$$

The actual comparison is performed for each dataset independently according to the average error $\bar{e}_m$ as defined by (2) where $\mathbb{D}_m, \mathbb{G}_m$ are the disparity and GT values selected by a given mask $m \in \{$ *"shape", "box", "rand", "all"*$\}$.

$$\bar{e}_m = \frac{|\{\forall d_i \in \mathbb{D}_m, g_i \in \mathbb{G}_m : \neg correct(d_i, g_i)\}|}{|\{\forall g_i \in \mathbb{G}_m : g_i \in \mathbb{R}\}|} \quad (2)$$

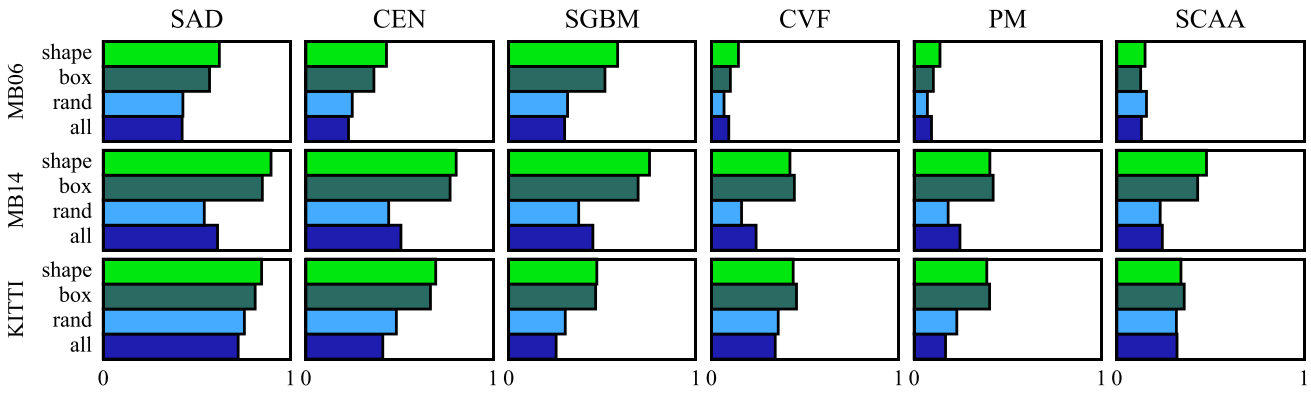

Fig. 7 Percentage of pixels with an error above 2px for all algorithms for the different masks and datasets (average error $\bar{e}$)

Figure 7 shows the result of the evaluation for all three datasets and all four mask types. The arithmetic average of the performance evaluated for 100 random masks are reported as *rand*. We chose to use a high threshold of 2pxl to distinguish the coarse cases "algorithm succeeded at finding a good correspondence" versus "algorithm could not determine a correct correspondence" as opposed to measuring small measurement errors. The performances of the different mask types creates a distinct picture. Section 6.2 will first interpret the results. The following Sect. 6.3 will then assign statistical significance to these interpretations.

6.2 Interpretation

The effect of applying the masks based on the identified hazards can be clearly seen. Table 6 summarizes the ratios between the error values of *shape* and *all*. The correctly masked areas (*shape*) have higher error ratios than the mean for the full image (*all*). The results for KITTI are much more erratic than the rest. The large amount of missing GT data in this dataset reduced its value for this evaluation drastically. The majority of *shape* mask areas have higher error ratios than the same-sized *box* mask areas. Newer and more complex algorithms generally score lower errors and have lower absolute differences between *shape* and *all* errors. There are two distinct groupings: *rand* masks have comparable results as *all* masks while *box* is comparable to *shape*. This suggests that box annotations can often be used instead of the time-consuming shape annotations. This allows for the following conclusions based on the different maskings: algorithms have higher error rates at annotated areas and score even higher error rates if the annotation's shape is preserved (*shape* vs. *box*). The effect of sampling patches of different sizes in each image is not prevalent (*rand* vs. *box*) and can be neglected.

Table 6 Ratio between average errors of *shape* and *all*: $\frac{\bar{e}_{\text{shape}}}{\bar{e}_{\text{all}}}$

Dataset	SAD	CEN	SGBM	CVF	PM	SCAA
MB06	1.47	1.89	1.95	1.55	1.49	1.15
MB14	1.47	1.58	1.67	1.76	1.65	1.97
KITTI	1.17	1.69	1.86	1.28	2.31	1.07

6.3 Statistical Significance

The intuitive grouping of the mask into groups (*all*, *rand*) and (*shape*, *box*) is now evaluated for its statistical significance. The null hypothesis H_0 we will test is that the average performance evaluated at two different mask-types is not distinguishable. More specifically, that the differences between pairings of measurements (x_i, y_i) are symmetrically distributed around zero. This hypothesis should be valid between the grouped mask types and invalid between the other types.

To test the hypothesis, parametric and non-parametric tests can be used. Parametric tests (e.g. T-test) need to make assumptions about the underlying distribution. Such assumptions would be detrimental for our analysis as they could introduce bias. From the possible non-parametric tests we chose the Wilcoxon signed rank test (Wilcoxon 1945) because of its robustness and the possibility to evaluate over all three datasets in one joined analysis (see Demšar 2006) for a comparison between similar suited tests). The evaluation of all three datasets in one test statistic increases the sampling size and, thus, the test's significance.

The Wilcoxon signed rank test works by calculating the absolute difference for each pair of measurements from the two distributions and sorting those differences in ascending order. The rank in this order is now summed up using the original sign of each of the differences and the absolute value of this sum is used as the test statistic W. Ties in the ranking receive all the same average over the tying ranks. The number of differences not equal to zero is denoted with N_r.

Distributions with a symmetry around zero will yield a sum that has an expected value of zero and a variance of

Table 7 Probability values z_W obtained from the Wilcoxon signed rank test for different pairings

Pairing	SAD	CEN	SGBM	CVF	PM	SCAA	Overall
all, rand	0.46	0.08	−0.33	−1.17	−0.81	−1.18	−1.10
shape, box	**−4.54**	**−3.22**	**−2.74**	−0.36	−0.20	−0.58	**−4.89**
all, shape	**4.79**	**5.95**	**4.89**	**2.19**	**4.03**	0.87	**9.64**
all, box	**4.00**	**5.72**	**4.92**	**2.50**	**4.01**	0.84	**9.43**
shape, rand	**−4.40**	**−4.98**	**−4.16**	**−2.36**	**−3.21**	−1.28	**−8.53**
box, rand	**−3.55**	**−4.70**	**−4.15**	**−2.48**	**−3.37**	−1.40	**−8.23**

Bold entries represent rejected null hypothis when using a typical significance level of 5% (translates to a z value of +/−1.96). Negative values mean the first entry of the pairing was more difficult that the second; positive values signify the opposite

$var_W = N_r(N_r+1)(2N_r+1)/6$. For $N_r > 9$ the distribution of W approaches a normal distribution with $\sigma_W = \sqrt{var_W}$ and $z_W = W/\sigma_W$. These resulting probability values z_W can be used as a measure for rejecting the null-hypothesis if z_W is larger than z_{Wc} based on the selected significance level.

In our case we calculate the differences using average performance between two mask variants for each single test case (stereo image pair) from the datasets and then sort all differences by their absolute value. The resulting sum of the signed ranks is divided by σ_W for the corresponding N_r of that comparison yielding a single z value each. This test is performed for all relevant pairings of masks and for each algorithm, but we will combine the differences for all datasets. Finally we also calculate the overall z value for each pairing by evaluating the cumulation of all algorithm results. Table 7 shows the summarized results for all tests. The 100 samples of each mask generated for *rand* are used to calculate 100 times the value of z_W for each combination that contains *rand*. The table entry contains the arithmetic average of all 100 values. For this evaluation we keep the sign of the resulting test statistic to preserve the direction of each comparison. The decision whether to accept or reject the null hypothesis (distribution of results from different masks are the same) is based on the selected significance level. This percentage describes the probability of rejecting a true null hypothesis (*type I error*). We now apply a significance level of 5% to the data which translates to a z value of $+/-1.96$. All null hypothesis with an absolute z_W value of higher than $z_{Wc} = 1.96$ can be rejected.

This results in the following observations:

- (*all*, *rand*) is not significantly different, the null-hypothesis that both confirm to the same distribution can be accepted
- (*shape*, *box*) is significantly different, *shape* is more difficult than *box*
- (*all*, *shape*) has the most significant difference, *shape* is much more difficult than *all*. The pairing *all*, *box* is also presenting the same level of significant differences. (*shape*, *rand*) and (*box*, *rand*) show slightly less significance but are still very definite: both *shape* and *box* are significantly more difficult than *rand*
- The significance of the results varies widely between the different algorithms. Older and real-time algorithms tend to show the highest test statistics. SCAA results in the same trends as the remaining algorithms but stays always below the significance level of 5%.

The evaluation paints a clear overall picture: areas identified by the CV experts as containing a visual hazard guided by the CV-HAZOP checklist are especially challenging for the selected CV algorithms. Focusing on these challenging areas is beneficial for robustness evaluations since it creates more meaningful test cases.

7 Conclusion

Many critical situations and relations have the potential to reduce the quality and functionality of CV systems. The creation of a comprehensive checklist containing these elements is a crucial component on the road towards systematic validation of CV algorithms. This paper presents the efforts of several experts from the fields of CV as well as risk and safety assessment to systematically create such a list. To the authors' best knowledge, this is the first time that the risk analysis method HAZOP has been applied extensively to the field of computer vision.

The CV-HAZOP is performed by first introducing a generic CV model which is based upon information flow and transformation. The model partitions the system into multiple subsystems which are called locations. A set of parameters for each location is defined, that characterize the location's individual influence on information. Additional special CV-relevant "guide words" are introduced that represent deviations of parameters with the potential to create hazards. The execution of the HAZOP was performed by a number of authors in parallel, assigning meanings to each combination of guide words and parameters to

identify hazards. The individual findings were discussed and merged into one resulting CV-HAZOP list. A guideline for using the hazard list as a tool for evaluating and improving the quality and thoroughness of test datasets is provided.

The CV-HAZOP has produced a comprehensive checklist of hazards for the generic CV algorithm with over 900 unique entries. Each individual hazard is now referable by a unique hazard identifier (HID). It supports structured analysis of existing datasets and calculation of their hazard coverage in respect to the checklist. We present an example by applying the proposed guidelines to popular stereo vision datasets and finally evaluate the impact of identified hazards on stereo vision performance. The results show a clear correlation: identified hazards reduce output quality.

8 Outlook

The creation or combination and completion of test datasets using our checklist is the logical next step. We plan to guide the creation of a stereo vision test dataset with known coverage of hazards from our checklist. Another idea is the creation of test data that gradually increases the influence of specific hazards (e.g. amount of low contrast textures). This allows to find the point of failure and get an accurate estimation about the robustness of an algorithm when facing a specific hazard. The usage of our checklist can also be streamlined. Pre-filtered lists for common applications and domains provide specific lists without the need of manual adjustments. We are also investigating the automatic detection of hazards, i.e. algorithmic checks to determine if and where a hazard is present in a test image. This will reduce the manual task of categorizing test data and in the long run should lead to a fully automatic CV validation framework.

Our HAZOP checklist is not considered final. It will be updated to include lessons learned during evaluations and testing or even after tested systems are put into operation. By sharing this information with the community over our public HAZOP database we hope to increase quality and reduce effort in CV robustness evaluation. At this stage, the CV-HAZOP becomes a structured and accessible reference hub for sharing experiences with CV algorithm development, usage, and maintenance.

Acknowledgements Special thanks for their extensive CV-HAZOP contributions go to Lawitzky G., Wichert G., Feiten W. (Siemens Munich), Köthe U. (HCI Heidelberg), Fischer J. (Fraunhofer IPA), and Zinner C. (AIT). Thanks to Cho J.-H. (TU Wien) and Beham M. (AIT) for their help with the example chapter. The creation of the CV-HAZOP as well as this work have been funded by the ARTEMIS Project R3-COP, No. 100233 and the European Initiative to Enable Validation for Highly Automated Safe and Secure Systems (ENABLE-S3) Joint Undertaking under grant agreement grant agreement No. 692455. This joint undertaking receives support from the European Union's HORIZON 2020 research and innovation programme and Austria, Denmark, Germany, Finland, Czech Republic, Italy, Spain, Portugal, Poland, Ireland, Belgium, France, Netherlands, United Kingdom, Slovakia, Norway. Additional support was received by the project autoBAHN2020 funded by the Austrian Research Promotion Agency with Contract Number 848896.

References

Aloimonos, J. Y., & Shulman, D. (1989). *Integration of visual modules: An extension of the Marr paradigm*. Boston: Academic Press Professional Inc.

Bleyer, M., Rhemann, C., & Rother, C. (2011). Patchmatch stereo–stereo matching with slanted support windows. In *British machine vision conference*.

Bowyer, K., & Phillips, P. J. (1998). *Empirical evaluation techniques in computer vision*. Los Alamitos, CA: IEEE Computer Society Press.

Bowyer, K., Kranenburg, C., & Dougherty, S. (2001). Edge detector evaluation using empirical ROC curves. *Computer Vision and Image Understanding*, *84*(1), 77–103.

Center for Chemical Process Safety. (1992). *Guidelines for hazard evaluation procedures, with worked examples* (2nd ed.). Hoboken: Wiley.

Department of Defense. (1949). *Procedures for Performing a Failure Mode, Effects and Criticality Analysis, MIL-STD-1629A*.

Department of Defense. (1991). *Reliability Prediction of Electronic Equipment: MIL-HDBK-217F*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Fenelon, P., & Hebbron, B. (1994). Applying HAZOP to software engineering models. In *Risk management and critical protective systems: proceedings of SARSS* (pp. 11–116).

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381395.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Computer vision and pattern recognition*.

Goseva-Popstojanova, K., Hassan, A., Guedem, A., Abdelmoez, W., Nassar, D. E. M., Ammar, H., et al. (2003). Architectural-level risk analysis using UML. *IEEE Transactions on Software Engineering*, *29*(10), 946–960.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, *42*(6), 1887–1896.

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 328–341.

Honauer, K., Maier-Hein, L., & Kondermann, D. (2015). The HCI stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *The IEEE international conference on computer vision (ICCV)*.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Humenberger, M., Zinner, C., Weber, M., Kubinger, W., & Vincze, M. (2010). A fast stereo matching algorithm suitable for embedded real-time systems. *Computer Vision and Image Understanding*, *114*, 1180–1202.

International Electrotechnical Commission. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems—part 4: Definitions and abbreviations: IEC 61508-4*.

Kadiofsky, T., Weichselbaum, J., & Zinner, C. (2012). Off-road terrain mapping based on dense hierarchical real-time stereo vision. In *Advances in visual computing*. Lecture Notes in Computer Science (Vol. 7431, pp. 404–415). Berlin: Springer.

Kajiya, J. T. (1986). The rendering equation. In *SIGGRAPH conference proceedings* (Vol. 20, No. 4, pp. 143–150).

Kletz, T. A. (1983). *HAZOP and HAZAN notes on the identification and assessment of hazards*. The Institution of Chemical Engineers.

Kondermann, D. (2013). Ground truth design principles: An overview. In Proceedings of the international workshop on video and image ground truth in computer vision applications, VIGTA '13 (pp. 5:1–5:4). ACM, New York, NY, USA

Kondermann, D., Nair, R., Meister, S., Mischler, W., Güssefeld, B., Honauer, K., et al. (2015). Stereo ground truth with error bars. In *Asian conference on computer vision*.

Konolige, K. (1998). Small vision systems: Hardware and implementation. In *Robotics research*. Berlin: Springer.

Laprie, J. (1992). Dependability: Basic concepts and terminology. In *Dependable computing and fault-tolerant systems* (Vol. 5). Berlin: Springer.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Matthias, B., Oberer-Treitz, S., Staab, H., Schuller, E., & Peldschus, S. (2010). Injury risk quantification for industrial robots in collaborative operation with humans. In *Proceedings of the of 41st international symposium on robotics and 6th German conference on robotics*.

Mei, X., Sun, X., Dong, W., Wang, H., & Zhang, X. (2013). Segment-tree based cost aggregation for stereo matching. In *Computer vision and pattern recognition* (pp. 313–320).

Min, J., Powell, M., & Bowyer, K. W. (2004). Automated performance evaluation of range image segmentation algorithms. *IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics*, *34*, 263–271.

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, *47*(2), 617–644.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLOS Computational Biology*, *4*, e27.

Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, et al. (2006). Dataset issues in object recognition. In *Toward category-level object recognition* (pp. 29–48). Springer.

Rhemann, C., Hosni, A., Bleyer, M., Rother, C., & Gelautz, M. (2011). Fast cost-volume filtering for visual correspondence and beyond. In *Computer vision and pattern recognition* (pp. 3017–3024).

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1), 7–42.

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *Computer vision and pattern recognition*.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X., & Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern recognition* (pp. 31–42). Springer.

Schlick, R., Herzner, W., & Jöbstl, E. (2011). Fault-based generation of test cases from UML-models approach and some experiences. In *Computer safety, reliability, and security*. Lecture Notes in Computer Science (Vol. 6894, pp. 270–283). Berlin: Springer.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, *37*(1), 10–21.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Champaign: University of Illinois Press.

Strecha, C., von Hansen, W., Van Gool, L., Fua, P., & Thoennessen, U. (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer vision and pattern recognition*.

Takanen, A., DeMott, J., & Miller, C. (2008). *Fuzzing for software security testing and quality assurance*. Artech House on Demand.

Thacker, N., Clark, A., Barron, J., Ross Beveridge, J., Courtney, P., Crum, W., et al. (2008). Performance characterization in computer vision: A guide to best practices. *Computer Vision and Image Understanding*, *109*(3), 305–334.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Computer vision and pattern recognition* (pp. 1521–1528).

Vesely, W. E., Goldberg, F. F., Roberts, N. H., & Haasl, D. F. (1981). Fault tree handbook. In *Systems and reliability research*. Office of Nuclear Regulatory Research: NRC.

Von Bertalanffy, L. (1968). *General systems theory*. New York 41973, 40.

Van der Spek, R., & Spijkervet, A. (1997). Knowledge management: Dealing intelligently with knowledge. In *Knowledge management and its integrative elements* (pp. 31–60).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Zhang, K., Fang, Y., Min, D., Sun, L., Yang, S., Yan, S., & Tian, Q. (2014). Cross-scale cost aggregation for stereo matching. In *Computer vision and pattern recognition*.

2.2 Analyzing Computer Vision Data - The Good, the Bad and the Ugly

Analyzing Computer Vision Data - The Good, the Bad and the Ugly

Oliver Zendel Katrin Honauer Markus Murschitz Martin Humenberger
Gustavo Fernández Domínguez
AIT, Austrian Institute of Technology, Donau-City-Strasse 1, 1220, Vienna, Austria
HCI, IWR at Heidelberg University, Berliner Strasse 43 D-69120 Heidelberg, Germany
{oliver.zendel;markus.murschitz;martin.humenberger;gustavojavier.fernandez}@ait.ac.at,
katrin.honauer@iwr.uni-heidelberg.de

Abstract

In recent years, a great number of datasets were published to train and evaluate computer vision (CV) algorithms. These valuable contributions helped to push CV solutions to a level where they can be used for safety-relevant applications, such as autonomous driving.

However, major questions concerning quality and usefulness of test data for CV evaluation are still unanswered. Researchers and engineers try to cover all test cases by using as much test data as possible.

In this paper, we propose a different solution for this challenge. We introduce a method for dataset analysis which builds upon an improved version of the CV-HAZOP checklist, a list of potential hazards within the CV domain. Picking stereo vision as an example, we provide an extensive survey of 28 datasets covering the last two decades. We create a tailored checklist and apply it to the datasets Middlebury, KITTI, Sintel, Freiburg, and HCI to present a thorough characterization and quantitative comparison. We confirm the usability of our checklist for identification of challenging stereo situations by applying nine state-of-the-art stereo matching algorithms on the analyzed datasets, showing that hazard frames correlate with difficult frames. We show that challenging datasets still allow a meaningful algorithm evaluation even for small subsets. Finally, we provide a list of missing test cases that are still not covered by current datasets as inspiration for researchers who want to participate in future dataset creation.

1. Introduction

Vision solutions are used in safety critical applications such as self-driving cars and guided surgical procedures. Rigorous quality assurance measures are thus needed to ensure safe operations. Software quality assurance provides two main techniques that can be applied in CV: verification and validation (V&V). Verification is the process of checking whether a given implementation fulfills the specifications used to define the program's behavior. In essence these are semi-automatic or automatic checks to detect software bugs and glitches. Validation on the other hand evaluates if the system fulfills a given task even under difficult circumstances. This is done by using test datasets and comparing the results obtained from the system to a defined ground truth (GT). Major questions about the quality and usefulness of test data for CV evaluation are still unanswered: What are the characteristics of a good dataset? How can shortcomings be identified and supplemented to create test datasets which are truly effective at uncovering algorithmic shortcomings? In this work we tackle the question: What constitutes good test data for robustness testing, *i.e.* the detection of possible shortcomings and weaknesses. We show that special care should be taken to cover a wide variety of difficult situations because whether for validation of CV algorithms or for training applications: Datasets need a mixture of positive cases (*the Good*), border cases (*the Bad*), and negative test cases (*the Ugly*). This paper focuses on test data for validating stereo vision algorithms but the presented methodology is applicable to basically all CV algorithms as well as the composition of machine learning training data.

To give an idea about the impact of selected datasets, Figure 1 shows the number of papers which cite stereo vision datasets published annually at three major computer vision conferences (CVPR, ICCV, and ECCV). It is interesting to note that the popular Middlebury dataset (indoor scenes) was recently overtaken by KITTI (driving scenes) which shows the importance of stereo vision in the field of autonomous driving and driver assistance systems.

Section 2 gives a thorough overview and listing of 28 stereo vision datasets and summarizes how content has changed historically. Section 3.1 reviews CV-HAZOP, a tool for systematic analysis of test datasets. It presents our improvements on the method: specialization of generic

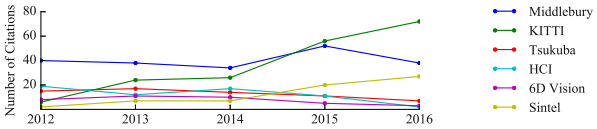

Figure 1. Number of stereo dataset citations published at CVPR+ICCV+ECCV for the years 2012-2016.

entries and instructions for easier analysis using the checklists. We apply the proposed concepts and create a specific checklist of dangerous/risky situations for stereo vision in Section 4.1. We evaluate five representative stereo vision datasets by using the proposed methodology in Section 4.2. In addition a range of stereo vision algorithms is evaluated in Section 4.3 using both traditional metrics and new metrics based on the results obtained by our checklist. Section 4.4 shows that the usage of challenging frames results in a comparable overall outcome even for a small number of test cases. Our checklist contains many critical situations that have not been found in any of the datasets. Section 4.5 presents this useful information for designing future datasets while the lessons-learned are shown in Section 4.6. Finally, Section 5 summarizes all findings and contributions of this paper.

2. State-of-the-Art

Reference data is the basis for performance analysis in computer vision. High-quality data is always well received in the community because it is essential to evaluate algorithm performance allowing the development of more accurate algorithms. Moreover, an objective comparison between algorithms using standardized data is important for a practical understanding of the current state-of-the-art in the respective area. Progress in stereo algorithm performance and the emerging applications of stereo technology motivate the need for more challenging datasets with accurate GT which emerges as a field of research. Among many others, examples of application domains are: autonomous driving (AD) [42, 66, 25, 23, 41, 60], space [24], agriculture [46], and medicine [6, 37, 36]. Early research introduced first datasets and performance metrics to show comparable results on the proposed algorithms. Initially, no common sequences/datasets were adopted. A clear domain or standard performance metrics definition were missing as well. Through the years, the CV community realized that thorough performance evaluation opens many research possibilities such as introduction of new datasets covering different scenarios and situations, analysis of performance metrics or online benchmarks comparing different algorithms. We now present the evolution of stereo vision datasets by comparing 28 datasets of the last two decades[1]. Table 1 gives an overview and presents quantitative characteristics of each dataset while Figure 2 shows representative images. We are focusing on the stereo vision test data. Many datasets contain additional GT (*e.g.* flow, segmentation, instances).

We will not compare datasets that have only RGBD data (no second camera image, *e.g.* NYU RGB-D [63, 44], TUM RGB-D [67] or the Berkeley dataset [22]). Please refer to the recent work of Firman [13] instead. There have been previous surveys on stereo vision and the interested reader is referred to [33, 57, 4, 32, 62, 30, 19].

2.1. Dataset Survey

In 2002 the Middlebury group proposed a taxonomy and a comparison framework of two-frame stereo correspondence algorithms [57]. The Middlebury website [68] evaluates stereo algorithms online, reports the performance of submitted algorithms, and offers stereo correspondence software for download. Over the years, the datasets were regularly updated: 6 datasets of piecewise planar scenes (2001), 32 datasets using structured light (between 2003 and 2006) and 43 high-resolution datasets with subpixel accurate ground truth (2014). EISATS [52] provides different video sequences for the purpose of performance evaluation. Traffic scenario scenes for evaluation of motion analysis, stereo vision, and optical flow algorithms are available to the community. Stereo sequences cover: Night vision (S1), synthesized (S2), color (S3), gray-level (S4&6), trinocular (S5&9), and consecutive stereo image pairs (S7). Neilson and Yang [45] introduced synthetic stereo pairs which were used to show their new evaluation method named cluster ranking. The dataset consists of 30 different stereo pairs containing three different baseline separations and three different noise levels and includes disparity maps and evaluation masks [48]. New College [65] is a large dataset ($\sim$ 30 GB) collected through the parks and campus of Oxford New College. The dataset focuses on outdoor SLAM (Simultaneous Localization and Mapping) applications and includes trajectories, stereo/omnidirectional imagery, as well as laser range/reflectance data. Pittsburgh Fast-Food [8] is a dataset containing 61 categories of food items. It aims to provide standard baselines for evaluating the accuracy of CV algorithms. EVD [9] dataset was developed for evaluating MODS (Matching On Demand with view Synthesis), an algorithm for wide-baseline matching of outdoor scenes but only includes homography data as GT. Ford Campus [50] dataset ($\sim$100 GB) is recorded using a 3D scanner laser and an omnidirectional camera intended for testing SLAM algorithms for AD. In 2012 Geiger et al. [15] introduced the KITTI Vision Benchmark Suite

[1] We tried to include every stereo vision dataset that also publishes GT; some datasets without GT were added due to their popularity.

Table 1. Summary of stereo datasets. 'w. GT' = number of images available with GT data, 'wo' = number without GT data, 'GT-Acc.' = GT accuracy in pixels, † =GT reported but dense GT is not available or the GT is very sparse/semantically oriented) * = algorithm results offered as GT, <1/N = granularity better than 1/N, S = synthetic, R = real, 1 = single shots, 2 = sequences of length 2, N = longer sequences

NAME	YEAR	IMAGES				DESCRIPTION
		Resolution	w. GT / wo	GT-Acc.	Type	
Middlebury [57]	2002	410 x 370	6 / —	1/8	R1	Piecewise planar cardboards
Middlebury [58]	2003	410 x 370	2 / —	1/4	R1	Cluttered still life
Middlebury [21]	2007	1390 x 1110	27 / 3	1	R1	Cluttered still life
EISATS S1 [70]	2008	640 x 481	— / 1900	—	RN	Traffic scenes
EISATS S2 [71]	2008	640 x 480	498 / —	<1/256	SN	Traffic scenes
Neilson [45]	2008	400 x 400	270 / —	1/16	S1	Still scene with var. textures/noise
EISATS S6 [53]	2009	640 x 480	— / 177	—	RN	Traffic scenes
New College [65]	2009	512 x 384	— / >100000	—	RN	Outdoor scenes for SLAM
Pittsburgh [8]	2009	1024 x 768	— / 130	*	R1	Fast food items (61 categories)
EVD [9]	2011	1000 x 750	— / 15	—	R1	Wide baseline still lifes
Ford Campus [50]	2011	1024 x 768	— / >100000	—	RN	SLAM, dynamic environments
HCI-Robust [27]	2012	656 x 541	— / 462	—	RN	Difficult road scenes
KITTI 2012 [15]	2012	1226 x 224	194 / 195	1/256	R2	Suburbs w. little traffic day time
Leuven [31]	2012	316 x 25	20 / 50	†	RN	Traffic day time
Tsukuba [38]	2012	640 x 480	1800 / —	<1/256	SN	Office cubicle still life
HCI-Synth [17]	2013	960 x 540	12 / —	1/256	S1	Texture challenges
Stixel [51]	2013	1024 x 333	2988 / —	†	RN	Highway w. good/bad weather
Daimler Urban [59]	2014	1024 x 440	— / 70000	—	RN	Urban city scenes
Malaga Urban [2]	2014	1024 x 768	— / >100000	*	RN	Dynamic environments real traffic
Middlebury [56]	2014	1328 x 1108	28 / 15	<1/256	R1	Cluttered indoor still life
Cityscapes [10]	2015	2048 x 1024	— / 20000	*	R1	Urban scenes daytime
KITTI 2015 [40]	2015	1242 x 375	200 / 200	1/256	R2	Road scenes with traffic
MPI Sintel [5]	2015	1024 x 436	1064 / —	<1/256	SN	Adventure movie scenes
Freiburg CNN [47]	2016	960 x 540	35454 / —	<1/256	SN	Road scene, animation movie
HCI Training [26]	2016	2560 x 1080	1023 / —	<1/256	RN	Difficult road scenes
SYNTHIA [55]	2016	960 x 720	>100000 / —	<1/256	SN	Diverse driving scenes
Virtual KITTI [14]	2016	1242 x 375	2126 / —	<1/256	SN	Suburban roads, currently RGBD
Oxford RobotCar [35]	To appear	1280 x 960	>100000 / —	<1/256	RN	Driving under varying weather and seasons

Figure 2. Excerpts from the discussed datasets. Images taken from the sources described in Table 1.

which includes a number of benchmarks. Stereo and optical flow data for close to 200 frames are provided. In addition, annotations include semantic and instance labels and longer image sequences of 20 frames per scene and there are about 200 frames where GT is withheld to ensure a fair evaluation on their website. In 2015 an updated

version of the dataset was released containing 400 image pairs of dynamic city scenes (200 for training and 200 for testing) and GT which was semi-automatically generated. Pixels are correctly estimated if the disparity or flow endpoint error is below a certain threshold, either 3 pixels or 5%, and it is required that the methods use the same parameter set for all test pairs. Their focus is on AD with the aim to reduce bias between real data and data generated under controlled conditions, *i.e.* laboratory environments. Objects such as cars and people are visible on each image. The Leuven [31] dataset presents image pairs from two cameras separated 1.5 meter apart from each other. The data was acquired in a public urban environment and contains both object class segmentation and dense stereo reconstruction GT for real world data. Tsukuba [38] dataset is a synthetic photo-realistic video dataset created as an reenactment of their well-known *head and lamp* stereo scene [43]. They include computer generated GT data for parameters, measurements, 3D position and distances. The 6D Vision group [11] makes two different datasets available to the community. The Daimler Urban Dataset [59] consists of video sequences recorded in urban traffic. Five semantic classes are defined (*building, ground, pedestrian, sky, and vehicle*) and 10% of the dataset is pixel-annotated using these classes. The Stixel Dataset [51] consists of 12 annotated stereo sequences acquired on a highway. Vehicle data, camera calibration, and GT generated by a fusing informations from manual annotations with ego-motion estimations are provided. HCI-Synth [17] contains four datasets, each covering a specific issue in stereo vision: visual artifacts, foreground fattening, decalibration, and textureless areas. Malaga Urban dataset [2] was recorded in urban scenarios using 9 cameras and 5 laser scanners containing real-life traffic scenes. The dataset is oriented toward object detection, SLAM, and visual odometry algorithms. The Cityscapes Dataset [10] was gathered entirely in urban street scenes focusing on semantic urban scene understanding. The dataset was recorded across several cities and different seasons. A benchmark suite, an evaluation server, and annotations (detailed for 5000 images and coarse for 20000) are also provided. The MPI Sintel Dataset [5] is derived from the animated short film Sintel containing diverse effects such as scene structure, blur, different illumination, and atmospheric effects. It is designed for the evaluation of optical flow, segmentation and stereo vision. Virtual KITTI [14] is a synthetic video dataset generated using virtual worlds. The scenarios comprise urban settings and the dataset is focused on multi-object tracking. No stereo setup has been released at the time of writing this paper (only RGBD). SYNTHIA (SYNTHetic collection of Imagery and Annotations) [55] is a synthetic dataset collected using 8 RGB cameras and 8 depth sensors. The data was acquired in different scenarios (cities, highways and green areas) under different illumination and weather conditions. The Oxford RobotCar Dataset [35] was collected by driving over the same route in Oxford throughout the year and thus represents good variations in seasons and weather.

2.2. Toward Optimal Test Data

The core problem of test data design is choosing the right number and kind of test cases. Some works in the CV community increased the number of sequences to the hundreds [12, 64, 34], but using more sequences does not necessarily increase diversity or coverage. Besides that, more data requires more GT, and GT acquisition is well known for being an error-prone and tedious task. Many recent works generate synthetic test data, where GT generation is more feasible and accuracy is higher (see [55, 18, 17, 49, 1, 5]). Another problem is dataset bias: test datasets without enough variation cannot reflect real world performance. Thus, researchers have begun to assess the role of diversity, coverage, and dataset bias. Torralba *et al.* [69] analyzed dataset bias, by training image classifiers to learn the dataset they belong to. The VOT challenge [29] performs clustering of a huge pool of sequences to reduce the size of the dataset to be evaluated while keeping in mind the diversity of the selected data. Zendel et al. [74] use a risk analysis procedure called Hazard and Operability Study (HAZOP) to evaluate and improve test datasets. HAZOP identifies difficult situations and aspects present in the dataset showing the hazard coverage of the dataset.

There are three main categories of test cases in traditional software quality assurance: positive test cases, border cases, and negative test cases. Positive test cases [61] represent normality and shall pose no problem to the algorithm. Border cases [7] are on the brink between specified and unspecified behavior but should still create meaningful outputs. Negative test cases [61] are expected to fail, but the error behavior should be well-defined (*e.g.* marking areas without meaningful values as invalid).

In this paper we concentrate on selecting challenging (*i.e.* border and negative) test cases in datasets to improve testing for robustness.

3. Methodology

Now we want to analyze some of the datasets presented in the previous section in depth and evaluate which hazards are tested by these datasets. We propose a new methodology based on an existing idea: Applying risk analysis to CV. First, this quality assurance approach is presented. Then, we extend the methodology. Finally, we apply this method to selected stereo vision datasets in Section 4.

3.1. CV-HAZOP

The systematic analysis of aspects that can influence the output performance and safety of a system is called a risk

analysis. Zendel *et al*. [74] apply a standard risk analysis called HAZOP to generic computer vision algorithms. First, they define an abstract CV model. Its components and their parameters create the basis for the HAZOP study. Then, modifier words called guide words are used to create entries representing deviations from the expected. These deviations are applied to each parameter and lead to a multitude of initial entries for the analysis. CV experts assign meanings, consequences and eventually hazards to each of these initial entries. The resulting list of identified vulnerabilities can be used to evaluate existing datasets and plan new ones. Each list entry can be referenced using its unique hazard identifier (HID). This approach allows qualitative and quantitative evaluation of datasets by identifying individual test cases that satisfy a stated checklist entry. However, there is a shortcoming with the proposed method: In order to have a unified generic checklist, each entry needs to be interpreted by the dataset analysts to their individual opinion. This results in a lot of ambiguity as different analysts might read and interpret the same entry in considerably different ways when applying it to the actual task at hand. Therefore we improve their work in the following aspects:

- Creation of specialized checklists specific to individual use cases instead of having each analyst start with the generic risk analysis lists (see Section 3.2).
- Methodology for analyzing datasets using the specialized checklist in Section 3.3.
- Application of the presented methods by creating a specialized checklist for stereo vision (Section 4.1).
- Analysis of popular stereo vision datasets using the specialized checklist presented in Section 4.3.

3.2. Checklist Specialization

The process starts with the publicly available generic CV-HAZOP checklist and transforms it into a specific one suitable for a particular domain and task:

- Decide for each entry in the list whether the hazards are relevant in the context of the actual task at hand.
- Create a single consensus summary for the entry. Write down as precisely as possible what is expected to be in a test image to fulfill the entry.
- Avoid duplicates and generate a concise list with a minimum of redundancy.

Experience has shown that the resulting list has to be revised after being used by the analysts for the first time. This resolves misunderstandings as well as annotation bias and allows to further remove redundancies.

3.3. How to Analyze a Dataset

The main goal of dataset analysis is usually to find at least one example test image for each checklist entry. This creates a rough estimate of the covered risks. First the analyst has to acquire a general overview of the dataset by noting regularities and reoccurring themes as well as special visually difficult situations such as: light sources (l.s.) visible within the image, visible specular reflections of l.s., large glare spots, large reflections showing near-perfect mirroring, transparencies, overexposure, underexposure, and large occlusions.

Now the specialist tries to find a fitting test image for each entry in the list. The restrictions found at the description are mandatory and reflect the transition from a generic hazard to the specific one. The relevant image part reducing the output quality for the target application should be large enough to have a meaningful impact (*e.g*. 1/64 of the image) and there should be valid GT available at this location. Test cases fulfilling only a single hazard with no overlap are preferred if there are multiple candidates for one entry. Otherwise images having the strongest manifestation of the hazard with largest affected areas are chosen.

4. Results

The presented methodology is applied to the stereo vision use case. A specific checklist is created and used to analyze popular existing stereo vision datasets. A thorough evaluation over a wide range of stereo vision algorithms generates an appropriate background for the following test data analysis. We show correlations between difficulty of test cases and predefined hazards from the checklist, indicate remarks about dataset size, and close with an extensive list of open issues missed in current datasets.

4.1. Stereo Vision Checklist

For our stereo vision checklist we define this use case: Calculate disparity maps from two epipolar constrained images without the use of prior or subsequent frames. The domain for which the algorithms should work is selected with the test datasets in mind: indoor scenes and outdoor driving scenes. We exclude most temporal hazards but otherwise regard all generic entries as potential candidates for our stereo vision checklist. Thus, we start with about 750 generic entries. Many hazards can quickly be disregarded as being out-of-scope for stereo vision. The remaining 350 entries are discussed and specialized. During this process some entries are deemed to be too extreme for our domain and many entries result in duplicates which are already part of the new checklist. At the end we derive 117 specialized entries from the generic list. Table 2 shows an excerpt of representative entries from the full list[2]. Each example is later identified in at least one dataset during the analysis. See Figure 3 for examples to each entry.

[2] See supplemental material or vitro-testing.com for the full list.

Table 2. Excerpts from full list of hazards for stereo vision (simplified, l.s. = light source)

hid	Loc. / GW / Param.	meaning	entry
0	L.s. / No / Number	No l.s.	Highly underexposed image; only black-level noise
26	L. s. / Part of / Position	Part of l.s. is visible	L.s. in image is cut apart by image border
142	L. s. / Less / Beam prop.	Focused beam	Scene with half lit object leaving a large portion severely underexposed
183	Medium / Less / Transparency	Medium is optically thicker than expected	Fog or haze in image reduces visibility depending on distance from observer
376	Object / Less / Complexity	Object is less complex than expected	Scene contains simple object without texture or self-shading (*e.g.* grey opaque sphere)
476	Object / No / Reflectance	Obj. has no reflectance	Well-lit scene contains a very dark object without texture nor shading
482	Object / As well as / Reflectance	Obj. has both shiny and dull surface	Object has a large glare spot on its surface that obscures same areas in the left/right image
701	Objects / Spatial aper. / Reflectance	Refl. creates a chaotic pattern	Large parts of the image show an irregular distorted mirror-like reflection
904	Obs. / Faster / Position	Observer moves too fast	Image has parts with clearly visible motion blur
1090	Obs. / No / PSF	No optical blurring	Image contains strong aliasing artifacts

Figure 3. Identified hazards in datasets corresponding to Table 2

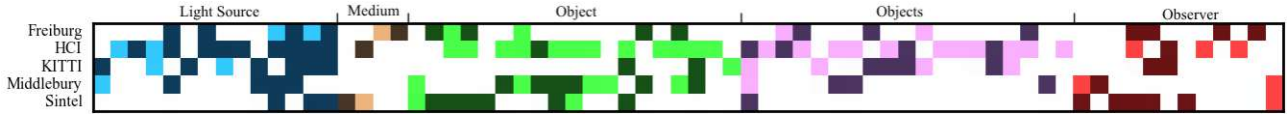

Figure 4. Distribution of hazards per dataset: Dark cells show identified hazards while light cells represent entries with no GT, too small area or disputed ones; color represents CV-HAZOP category.

4.2. Analyzing Test Data

Of all identified test datasets from Section 2 we concentrate on a specific subgroup: All datasets that are public, provide GT data, and have at least ten test images. This results in the following subsets: all Middlebury datasets, both KITTI datasets, Sintel, HCI Training 1K, and Freiburg[3]. The Oxford RobotCar and SYNTHIA datasets are certainly interesting for this evaluation but have been published too recently given their huge size for us to process.

The dataset analysis commences as described in Section 4.3. Two additional analysts as well as all authors participate, ensuring that each dataset is analyzed by at least two different people to reduce bias. In total, 76 hazards are found across all the datasets. They result in 48 unique hazards out of 117. Most hazards are found in the HCI Training Dataset, Freiburg, and Sintel (16 each) followed by the KITTI and Middlebury datasets (14 each). Figure 3 gives some examples of identified hazards. The entries correspond to the rows of Table 2. Some hazard entries are deemed to be unreliable for the upcoming evaluation due to missing GT, insufficient size, or disagreement between experts. These disputed entries were removed from the evaluation. Figure 4 visualizes the hazard distribution over all datasets. This still leaves 50 entries uncovered by any of the datasets. Section 4.5 will discuss these open issues.

[3]Freiburg is annotated without *flying things*. These scenes are too chaotic for analysts to evaluate in a reasonable time.

4.3. Dataset Evaluation

The following stereo vision algorithms are now evaluated on the analysed datasets: SAD + Texture Thresholding (TX) & Connected Component Filtering [28], SGM [20] with rank filtering (RSGM), Elas [16] + TX & Weighted Median Post Processing Filtering (WM), Cost-Volume Filtering (CVF) & WM[54], PatchMatch (PM) & WM [3], Cross-Scale Cost Aggregation using Census and Segment-Trees (ST) & WM [75, 39], SPSS [72], and MC-CNN [73] using their KITTI2012 pre-trained fast network. Average RMS and bad pixel scores for each test image in the datasets are calculated as evaluation metrics.

Figure 5 shows a summary of the difficulty for each dataset based on the performance of each algorithm. Unfilled bars visualize the relative amount of frames in the whole dataset with the specified difficulty, while filled bars denote the amount of hazard frames within this range of difficulty. All bars are normed to their respective maximum number.

As expected, the algorithms behave quite differently on the same test data due to their different implementations and performance varies depending on the dataset[4]. It is evident that hazard frames strongly group at the bins of higher difficulty. Filled bars are generally higher than unfilled bars for difficult frames (bins D/E) and lower for easier frames (bins A/B). This trend can be observed in each of the datasets for all algorithms.

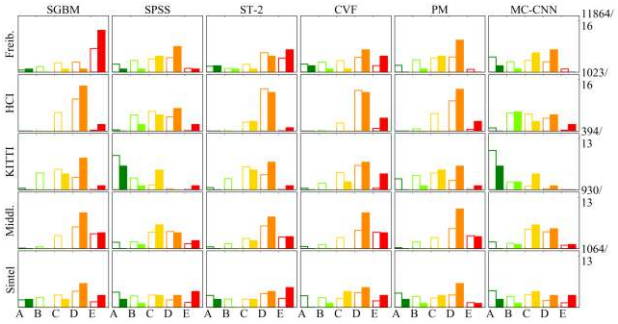

Figure 5. Difficulty distribution of frames in each dataset. Relative number of pixels having an error > 4 disparities sorted into 5 bins: A:[0-5%), B:[5-10%), C:[10-20%), D:[20-50%), E:[50-100%]. Right side: number of frames in full dataset (no-fill bars) / with hazards (solid bars). All bars (no-fill/solid) of a single plot add up to these respective numbers (first/second).

4.4. Data Size

One important aspect of test dataset design is using the right data size. Too much redundancy increases processing time and might drown relevant individual test cases in a flood of meaningless repetitions. Too few test cases, on the other hand, will prevent the detection of important shortcomings due to missing scenarios.

For our experiment we sort all frames by their difficulty according to performance per algorithms. We choose a subset of all frames and iteratively calculate the average performance over the subset adding easier frames with each step. In the first experiment we randomly pick frames from the dataset, achieving a good representation for the entire dataset. In our second experiment we only add the easiest frames of the dataset. In the third experiment we only use frames identified by the HAZOP analysis and add them in hardest-first manner. To make the results comparable we plot the accumulation of all frames up to the number of annotated hazard frames.

Figure 6 shows a comparison of the results (random, best first, HAZOP) for the Sintel dataset. Using only hazard frames allows the same level of distinction between algorithms with comparable numbers of images. Selecting hard frames is a valid way to evaluate algorithms. The advantage of using hazard frames in comparison to random picking is that they also give insights into why a specific test case failed.

[4]See supplemental material for addition algorithm performance graphs.

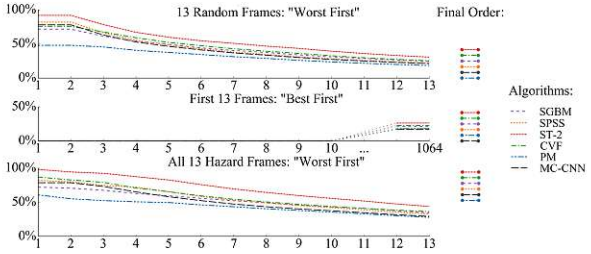

Figure 6. Comparison of cumulative average performance of 13 frames from Sintel: Random picking, easiest frames, hazard frames (all sorted by difficulty) using the bad pixel metric with a threshold of 4.

4.5. Missing Tests

There were numerous hazard entries which were not found in any of the test datasets examined by the analysts (Table 3). These entries were categorized into two groups: border cases and negative test cases. The distinction between the two is sometimes dependent on the domain (*e.g.* not every implementation has to work with a large field-of-view (FOV) or when there is rain/snow in the scene). For this checklist we tried to cover a very broad domain and require a lot of robustness from the algorithm, *i.e.* indoor scenes and outdoor street environments under difficult weather conditions. Using these guidelines we also decided on the clustering into *the Bad* and *the Ugly* groups. Positive test cases are usually easy to define. Therefore, we focus on difficult test cases.

4.6. Future Work

Testing algorithms with single test cases for each hazard allows for valuable insights, but more than a single data point is needed for representative statistics. Systematic test data, gradually increasing in difficulty, should be used to evaluate the breaking point of the algorithm (in regard to a specific hazard). Frame-based annotation should be augmented using labels within the images. This allows evaluations of hazards affecting smaller areas which otherwise get outweighed by the surrounding area's influences.

Focusing on the most difficult frames of a dataset can also give good indications about hazards without the need to inspect each frame. However, this can introduce a huge bias toward the evaluation metric used and propagate existing redundancy.

5. Conclusion

This paper focuses on analyzing datasets for their ability to test the robustness of CV applications. A thorough survey of 28 existing stereo vision test datasets demonstrates their

Table 3. Selection of hazards missing from current test datasets, see supplemental material for the full list

hid	entry
Border cases (*the Bad*)	
6	L.s. and its reflection are visible on the same epipolar line
12	Multiple l.s. are periodically placed and aligned on the same epipolar line
63	L.s. visible in image with a long elongated thin shape (*e.g.* neon tube) creating an unusual overexposed area
107	L.s. projects structured pattern onto a surface that produces two distinctly different Moire patterns in both images
259	Scene is split into two equal parts: one without particles and another with considerable amount of particles
310	Two different sized objects are positioned on the same epipolar line but their projected views are identical
341	Scene contains an expanding/shrinking object resulting in noticeable radial motion blur
479	Object has strongly reflecting material that mirrors larger parts found on the same epipolar line
523	Two partially transparent objects are entangled in such a way that both allow the view on each other
694	Scene contains a clear reflection of observer together with potential matching parts on the same epipolar
754	Scene contains a prominent rainbow effect (*i.e.* mist/haze with a view-depended colour band)
758	Scene contains pronounced refraction rings (*e.g.* oil slick)
803	Cameras have both a wide FOV (>135deg)
918	Lens body/lens hood is prolonged and its corners are thus blocking the view
926	Two cameras both have considerable comparable amount of dirt/pollution but with different distributions
1091	Very different textures in left and right image due to large scale Moire effects
Negative test cases (*the Ugly*)	
245	Cloud of visible particles (*e.g.* pollen, small leaves) in the air are obscuring the whole scene
504	Highly transparent object encompassing a second opaque object that gets distorted due to the other object's shape
695	Scene contains a large concave mirror that shows an clean upside-down copy of parts of the scenery
719	Observer is placed between two parallel mirrors facing each other so that "infinite" number of reflections occur
790	Left and right image are the same while showing a diverse scene
916	One camera lens contains dust/dried mud that creates a partially defocused area in the image
921	Lens is broken cleanly leaving a visible crack in the image's center
933	Images contain rolling shutter artifacts
955	Images contain considerable chromatic aberration and many visible edges
983	Images have considerable amounts of vignetting and scene contains many objects close to the observer
1094	One of the two sensors is somewhat out of focus
1105	Inter-lens reflections create visible copy of objects in the image
1162	Image before rectification originates from considerably rectangular pixels (instead of square, near to e.g. 2:1 ratio)
1166	Images contain strong static image noise for well-lit scenes
1261	One camera delivers negative image (or color channels swapped)
1265	Images use logarithmic quantization instead of linear or wrong gamma mapping

progression over time. We present an improved methodology based on the CV-HAZOP checklist analysis method that identifies challenging elements in datasets. We apply this methodology to selected popular stereo datasets to identify challenging test cases. Then, we evaluate a broad range of algorithms on those selected datasets. The correlation between frames identified as challenging and test case difficulty allows these conclusions: (i) cases marked as challenging are evidently difficult independent of dataset or algorithm choice, and (ii) challenging cases of a dataset are a representative subset of the entire dataset. Testing with challenging cases only yields similar results compared to the entire dataset but contains all listed challenges.

Most importantly, we present a list of challenges that are missing from all the selected datasets. This results in a roadmap of 32 practical inputs for researchers designing new datasets.

In our opinion, new datasets should increase difficulty and variability but not necessarily size: In addition to the easy cases (*the Good*), more border cases (*the Bad*) and negative test cases (*the Ugly*) should be added. Ultimately, this will increase applicability, usefulness, and the safety of CV solutions as well as systems that rely on them.

6. Acknowledgement

This project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 692480. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Saxony, Spain, Austria, Belgium, Slovakia. See www.iosense.eu; Thanks for proofreading and good suggestions go to Daniel Steininger (AIT) and Emma Alexander (Harvard).

References

[1] D. Biedermann, M. Ochs, and R. Mester. Evaluating visual ADAS components on the COnGRATS dataset. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016. 4

[2] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez. The málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario. *International Journal of Robotics Research*, 33(2):207–214, 2014. 3, 4

[3] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, 2011. 6

[4] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003. 2

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 3, 4

[6] F. Campo, F. Ruiz, and A. Sappa. Multimodal stereo vision systems: 3d data extraction and algorithm evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):437–446, 2012. 2

[7] J. Cem Kaner. What is a good test case? *STAR East*, 2003. 4

[8] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. PFID: Pittsburgh fast-food image dataset. In *Proceedings of International Conference on Image Processing*, 2009. 2, 3

[9] K. Cordes, B. Rosenhahn, and J. Ostermann. Increasing the accuracy of feature evaluation benchmarks using differential evolution. In *IEEE Symposium on Differential Evolution (SDE)*, 2011. 2, 3

[10] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015. 3, 4

[11] Daimler Böblingen, 6D-Vision. `http://www.6d-vision.com`. Accessed: 2016-11-15. 4

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 4

[13] M. Firman. RGBD Datasets: Past, Present and Future. In *CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis*, 2016. 2

[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3, 4

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 2, 3

[16] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010. 6

[17] R. Haeusler and D. Kondermann. Synthesizing real world stereo challenges. In *German Conference on Pattern Recognition*, pages 164–173. Springer, 2013. 3, 4

[18] V. Haltakov, C. Unger, and S. Ilic. Framework for Generation of Synthetic Ground Truth Data for Driver Assistance Applications. In J. Weickert, M. Hein, and B. Schiele, editors, *Pattern Recognition*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013. 4

[19] R. Hamzah and H. Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016. 2

[20] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):328–341, 2008. 6

[21] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, pages 1–8. IEEE, 2007. 3

[22] A. Janoch. The berkeley 3d object dataset. Master's thesis, EECS Department, University of California, Berkeley, May 2012. 2

[23] C. G. Keller, M. Enzweiler, and D. M. Gavrila. A new benchmark for stereo-based pedestrian detection. In *IEEE Intelligent Vehicles Symposium*, pages 691–696, 2011. 2

[24] W. Kim, A. Ansar, R. Steele, and R. Steinke. Performance analysis and validation of a stereo vision system. In *IEEE International Conference on Systems, Man and Cybernetics*, 2005. 2

[25] R. Klette, N. Kugrer, T. Vaudrey, K. Pauwels, M. van Hulle, S. Morales, F. I. Kandil, R. Haeusler, N. Pugeault, C. Rabe, and M. Lappe. Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. In *IEEE Intelligent Vehicles Symposium*, pages 2012–2026, 2011. 2

[26] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gussefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016. 3

[27] D. Kondermann, A. Sellent, B. Jähne, and J. Wingbermühle. Robust Vision Challenge, 2012. 3

[28] K. Konolige. Small vision systems: Hardware and implementation. In *Robotics Research*. Springer, 1998. 6

[29] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, Accepted, 2016. 4

[30] D. Kumari and K. Kaur. A survey on stereo matching techniques for 3D vision in image processing. *International Journal on Engineering and Manufacturing*, 4:40–49, 2016. 2

[31] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision*, 2012. 3, 4

[32] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos. Review of stereo vision algorithms: From software to hardware. *International Journal of Optomechatronics*, 2:435–462, 2008. 2

[33] M. Lemmens. A survey on stereo matching techniques. In *ISPRS Congress, commission V*, pages 11–23, 1998. 2

[34] A. Li, M. Lin, Y. Wu, M.-H. Yang, and S. Yan. NUS-PRO: A new visual tracking challenge. *TPAMI*, 38(2):335–349, 2016. 4

[35] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, to appear. 3, 4

[36] L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N. Clancy, D. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Müller-Stich, D. Oladokun, S. Röhl, T. R. dos Santos, H.-P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, , and D. Stoyanov. Comparative validation of single-shot optical techniques for laparoscopic 3d surface reconstruction. *Transactions on Medical Imaging*, 33(10):1913–1930, 2014. 2

[37] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974–996, 2013. 2

[38] M. Martorell, A. Maki, S. Martull, Y. Ohkawa, and K. Fukui. Towards a simulation driven stereo vision system. In *ICPR*, pages 1038–1042, 2012. 3, 4

[39] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition*, pages 313–320, 2013. 6

[40] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[41] N. Morales, G. Camellini, M. Felisa, P. Grisleri, and P. Zani. Performance analysis of stereo reconstruction algorithms. In *ITSC*, pages 1298–1303, 2013. 2

[42] S. Morales, T. Vaudrey, and R. Klette. Robustness evaluation of stereo algorithms on long stereo sequences. In *Intelligent Vehicles Symposium*, pages 347–352, 2009. 2

[43] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo-occlusion patterns in camera matrix. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 371–378. IEEE, 1996. 4

[44] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[45] D. Neilson and Y.-H. Yang. Evaluation of constructable match cost measures for stereo correspondence using cluster ranking. In *Computer Vision and Pattern Recognition, 2008. Proceedings CVPR'08, 2008 IEEE Computer Society Conference on*. IEEE, 2008. 2, 3

[46] M. Nielsen, H. Andersen, D. Slaughter, and E. Granum. High-accuracy stereo depth maps using structured light. *Precision Agriculture*, 8(49):49–62, 2007. 2

[47] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 3

[48] U. of Alberta Stereo Vision Research, 2010. 2

[49] N. Onkarappa and D. Sappa, A. Synthetic sequences and ground-truth flow field generation for algorithm validation. *Multimedia Tools and Applications*, 2013. 4

[50] G. Pandey, J. R. McBride, and R. M. Eustice. Ford campus vision and lidar data set. *International Journal of Robotics Research*, 30(13):1543–1552, 2011. 2, 3

[51] D. Pfeiffer, S. K. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *CVPR*, 2013. 3, 4

[52] Reinhard Klette: EISATS. http://ccv.wordpress.fos.auckland.ac.nz/eisats/. Accessed: 2016-11-15. 2

[53] R. Reulke, A. Luber, M. Haberjahn, and B. Piltz. Validierung von mobilen stereokamerasystemen in einem 3d-testfeld. (EPFL-CONF-155479), 2009. 3

[54] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Computer Vision and Pattern Recognition*, pages 3017–3024, 2011. 6

[55] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3, 4

[56] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer, 2014. 3

[57] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1/2/3):7–42, 2002. 2, 3

[58] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, pages 195–202. IEEE Computer Society, 2003. 3

[59] T. Scharwchter, M. Enzweiler, S. Roth, and U. Franke. Stixmantics: A medium-level model for real-time semantic scene understanding. In *ECCV*, 2014. 3, 4

[60] K. Schauwecker, S. Morales, S. Hermann, and R. Klette. A new benchmark for stereo-based pedestrian detection. In *IEEE Intelligent Vehicles Symposium*, 2011. 2

[61] G. S. Semwezi. Automation of negative testing. 2012. 4

[62] P. Sharma and N. Chitaliya. Obstacle avoidance using stereo vision: A survey. In *International Journal of Innovative Research in Computer and Communication Engineering*, pages 24–29, 2015. 2

[63] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011. 2

[64] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *TPAMI*, 36(7):1442–1468, 2014. 4

[65] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599, May 2009. 2, 3

[66] P. Steingrube, S. Gehrig, and U. Franke. *Performance evaluation of stereo algorithms for automotive applications*. Computer Vision Systems, 2009. 2

[67] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2

[68] The Middlebury Computer Vision Pages. http://vision.middlebury.edu/. Accessed: 2016-11-15. 2

[69] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011. 4

[70] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behavior on synthetic and real-world stereo sequences. In *IVCNZ*, pages 1–6, 2008. 3

[71] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *10th European Conference on Computer Vision (ECCV '08)*, pages 739–751, Berlin, Heidelberg, 2008. Springer-Verlag. 3

[72] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014. 6

[73] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 6

[74] O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner. CV-HAZOP: Introducing test data validation for computer vision. In *ICCV*, 2015. 4, 5

[75] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, S. Yan, and Q. Tian. Cross-scale cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition*, 2014. 6

2.3 WildDash - Creating Hazard-Aware Benchmarks

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 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

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

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*.

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.

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

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 Section 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 Section 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. 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 (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 Figure 1 for example images containing these hazards.

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

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

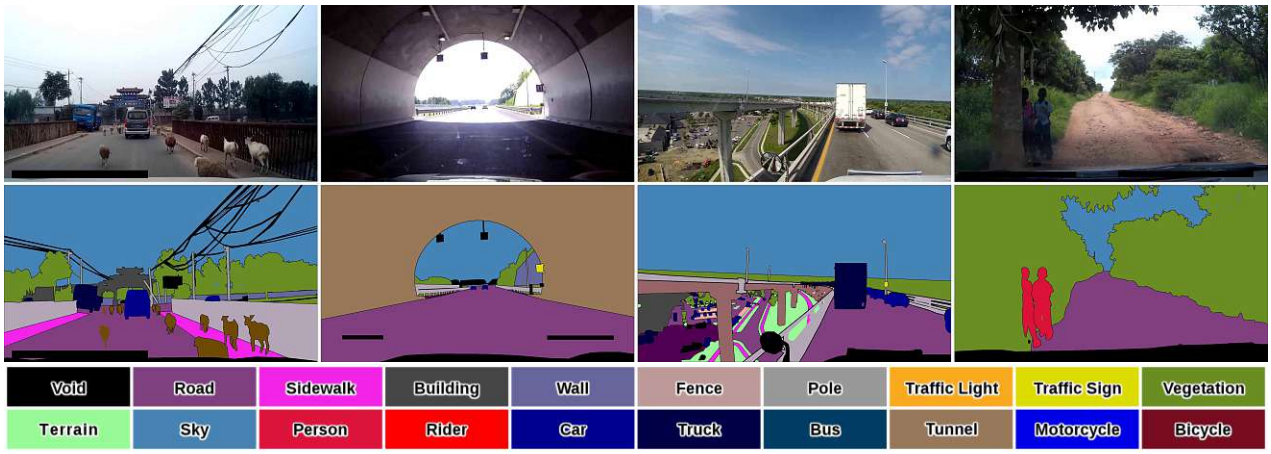

Figure 3. Positive test cases from wd_val_01 (cn0000, si0005, us0006, and zm0001) together with a visualization of the respective semantic segmentation and color legend.

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 (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'}.
- 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 Figure 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.

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

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		

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

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 x 896 pixels. As a pre-processing step, the *Mapillary Vistas* dataset has been rescaled and cropped to fit the resolution of Cityscapes (2048 x 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 *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.

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**

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 Figure 5. As long as input

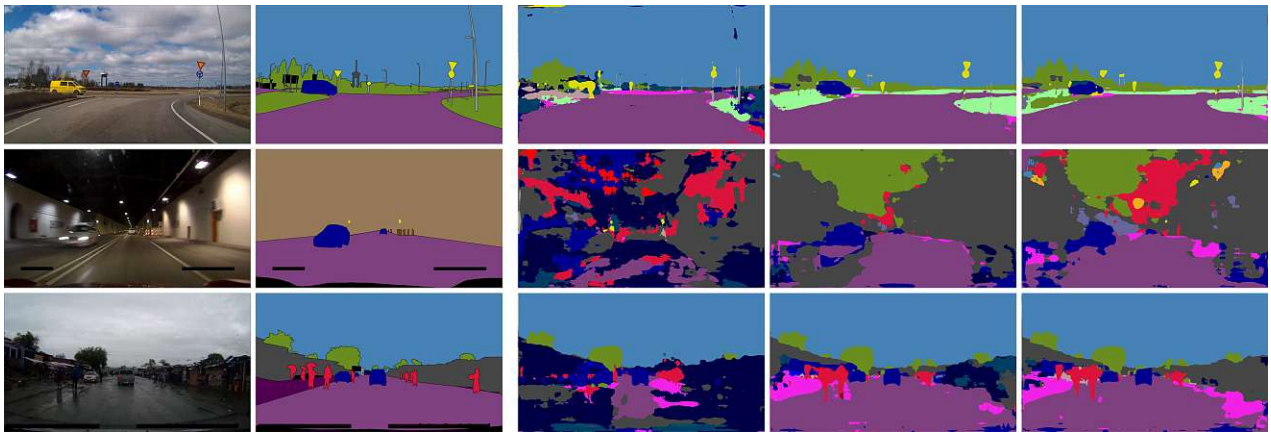

Figure 5. Qualitative results of our baseline models on *WildDash* validation images (left to right: input image, corresponding ground truth, and the inferred labelings of our baseline models cityscapes, mapillary, and mapillary+)

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, mod-

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 section 5, quantifies the potential negative influence of a specific hazard on the labeling performance

hazard	blur	cover-age	distor-tions	hood	occlu-sion	over-exp	under-exp	par-ticles	wind-screen	vari-ations
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

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

domain	city	high-way	off-road	over-land	sub-urban	tun-nel	curve	inter-section	round-about	stra-ight
mIoU	31.3	24.5	**32.7**	29.3	31.6	19.6	28.7	31.7	**36.6**	28.0

els 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. 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

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

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 for variations in road layouts, the best labeling scores are achieved in round-about 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 par-

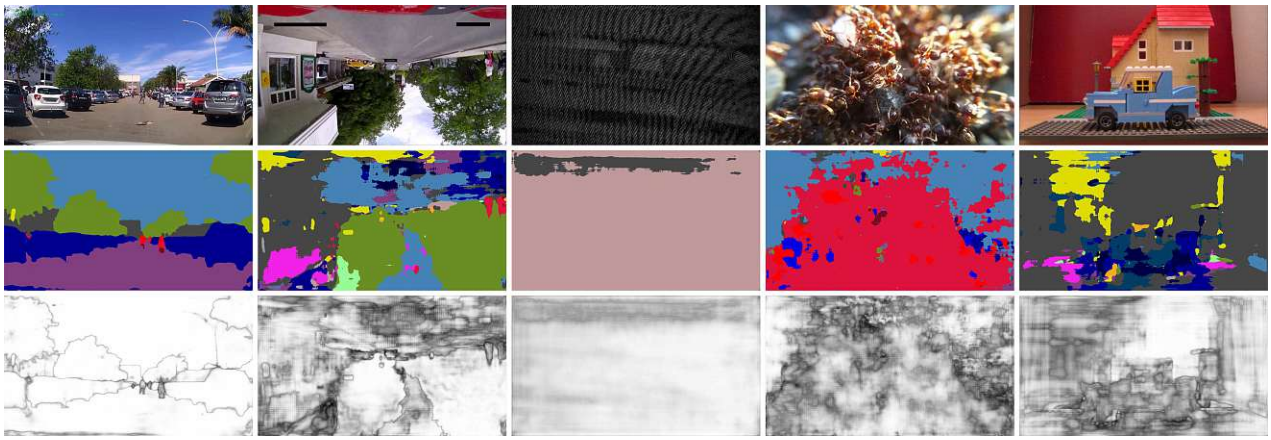

Figure 6. Input images, semantic segmentation results and corresponding confidence of baseline model mapillary+ on *WildDash* test images (left to right: positive test case, altered valid image, abstract image and two out-of-scope images).

tially 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.

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`.

9 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. (2011) 1521–1528
2. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: ECCV (1). (2008) 44–57
3. Franke, U., Gehrig, S., Rabe, C.: Daimler Böblingen, 6D-Vision. `http://www.6d-vision.com` Accessed: 2016-11-15.
4. Scharwächter, T., Enzweiler, M., Roth, S., Franke, U.: Stixmantics: A medium-level model for real-time semantic scene understanding. In: ECCV. (2014)
5. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: 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 Robotics and Automation Letters **2**(4) (2017) 2188–2193
7. Mighty AI: Mighty AI Sample Data. `https://info.mty.ai/semantic-segmentation-data` Accessed: 2018-03-07.
8. Mapillary Research: Mapillary Vistas Dataset. `https://www.mapillary.com/dataset/vistas` Accessed: 2018-02-16.
9. University of California, Berkeley, U.: Berkeley deep drive. `http://data-bdd.berkeley.edu/` Accessed: 2018-03-07.
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: 2018-02-16.
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., Schneider, L., Espinosa, A., Vazquez, D., Lopez, A.M., Franke, U., Pollefeys, M., Moure, J.C.: 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. (2017) 6670–6680
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. International Journal of Computer Vision **111**(1) (Jan 2015) 98–136
20. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: CVPR. (2017)

2.4 Unifying Panoptic Segmentation for Autonomous Driving

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

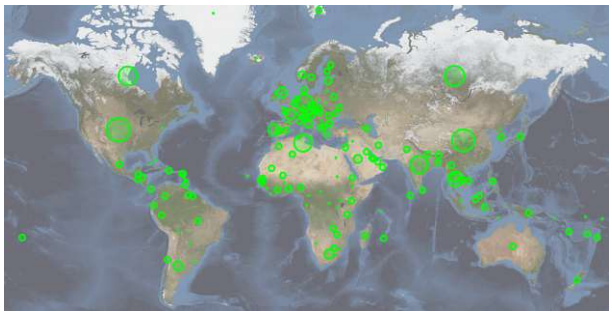

Figure 2. Visualization of Wilddash2 geographic distribution. Dots denote 1-9 scenes; small circles 10-50; medium circles: 50-200; large circles: >200 scenes. Globe courtesy of USGS [35].

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-

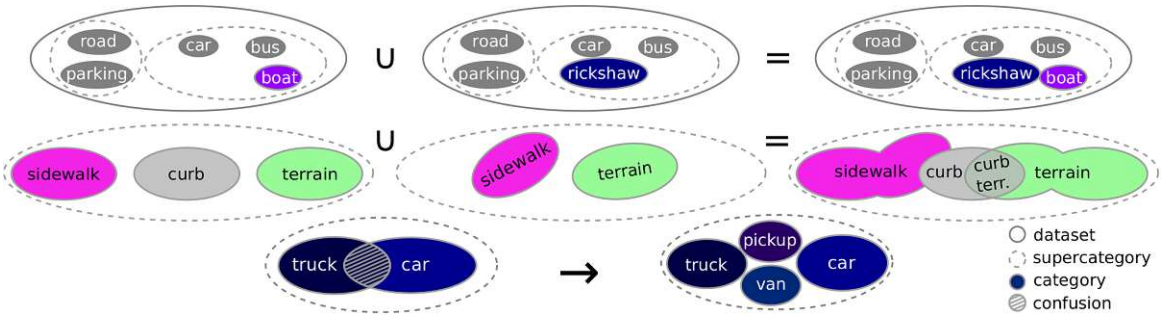

Figure 3. Conceptual depiction of label unification: (top) Organization and combination of disjunct categories and supercategories of two datasets. (center) merging and splitting of sets in case of label-policy-clashes of two datasets (see Figure 4). (bottom) cleaning up mixed categories by the introduction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

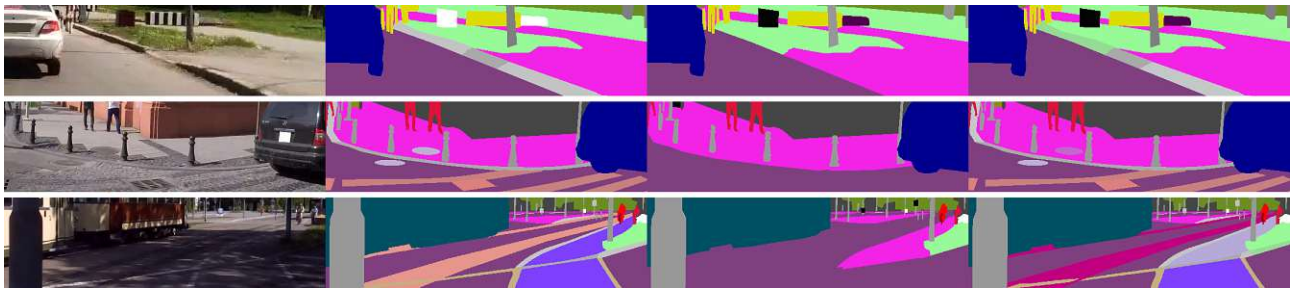

Figure 4. Example frames from WD2 visualizing the need for additional splitting of some labels. Left to right: crop from RGB image, GT using MVD classes, GT using Cityscapes classes, GT using WD2 classes. From top to bottom: ru0009_10000 (*curb* vs. *curb-terrain*), ga0004_10000 (*manhole* vs. *manhole-sidewalk*),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

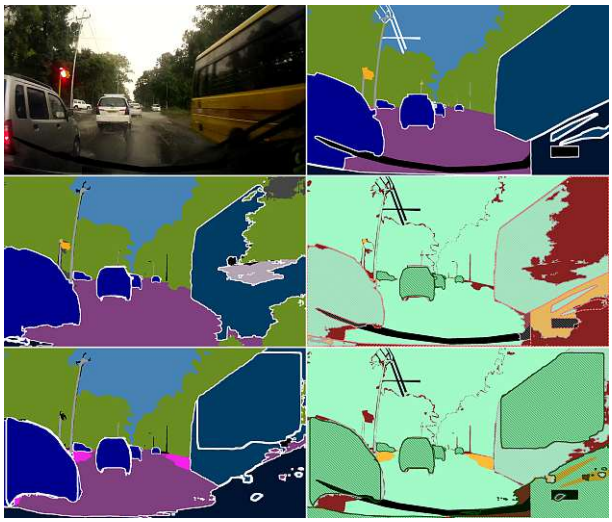

Figure 5. Visualization method for panoptic segmentation results. Top: WD2 scene in0090 RGB image and GT; Middle: result of the MVD-trained model (*mvd100*) and proposed difference image (see Section 4.3); Bottom: result of mixed MVD&WD2 model (*mix150*) and difference image (in0090 was part of the random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

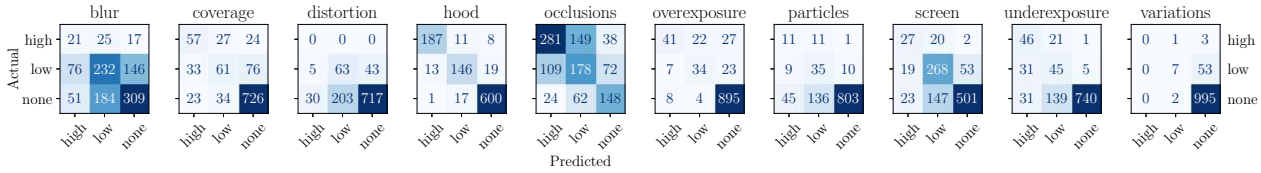

Figure 6. Confusion matrices for each prototype hazard classifiers.

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

CHAPTER 3

Conclusion

3.1 Summary

Computer vision applications rely heavily on image data for training and evaluations. Robustness of real-world solutions have to be tested with challenging data to identify potential problems during their development. How can we create difficult test cases or identify them in existing supplied datasets? This work presents a method for systematically listing circumstances and aspects which can have negative effects on their performance by applying risk analyis to computer vision applications. The method *Hazard and operability analysis* (**HAZOP**) is applied to create a list of generic **visual hazards**: CV-HAZOP. Evaluation of stereo vision datasets show, that areas with identified visual hazards lead to lower performances with a high statistical significance. The entries of the checklist indeed represent factors which can result in more **challenging test cases**. A specialized version of visual hazards for stereo vision is created to evaluate the difficulty of existing datasets. Five well-used datasets for benchmarking stereo vision are evaluated to identify test cases which contain specific visual hazards. The results are clear: existing benchmarking datasets contain very little challenging situations, but focus on mostly ideal conditions. Performances of multiple stereo vision algorithms are compared between frames identified as potentially difficult vs. easy. This confirms that the manual selection based on visual hazards has led to a more difficult subset. A new dataset with visual hazards in mind is created for semantic scene understanding: **Wilddash**. Scenes are added based on which specific hazards they contain. Algorithm performance for test case subsets containing certain hazards can be compared to the average performance to calculate hazard-specific performance impacts. Negative testing is introduced for segmentation tasks by comparing results for out-of-distribution test images (e.g. an aquarium scene for road scene understanding). A new evaluation metric expects invalidated results for negative tests while allowing an additional best-effort solution. The dataset and an automatic online evaluation platform for road scene understanding is

launched on **wilddash.cc**. The updated **Wilddash 2** (WD2) dataset is created, which increases in size by a factor of 20. It features a new **unified label policy** making it compatible with four established existing datasets for road scene understanding. The concepts of hazard-aware testing as well as negative testing are extended to panoptic segmentation and a new method for visualizing panoptic results is presented. Multiple panoptic segmentation prototypes are trained using various datasets with and without WD2 data. Adding WD2 data results in considerable gains when evaluating on WD2 benchmark data. No noticeable negative effects on their performance are introducing when testing on their respective original datasets. Lastly, a proof-of-concept for an automatic classifier to detect various visual hazards in image data is presented.

3.2 Outlook

The following topics and projects have started and will continue as a consequence of the presented work. They will provide continuity and long-lasting service to the scientific community to improve computer vision validation.

3.2.1 LiDAR and CV-HAZOP

The safety for many autonomous systems are increased by adding more types of sensors. Risks and performance degradation associated with one sensor can be mitigated by sensor fusion. LiDAR (light detection and ranging) sensors are quickly becoming a standard sensor for many applications. An ongoing risk-assessment and experiments will apply the risk analysis of CV-HAZOP to LiDAR sensors.

3.2.2 Wilddash Webservice

The public webservice launched together with the Wilddash datasets [ZHM+18, ZSR+22] continues to provide public leaderboards for semantic segmentation, instance segmenation, and panoptic segmentation. Ground-truth for the dedicated benchmarking dataset part of Wilddash is held secret to reduce potential dataset biases. It allows a fair comparison of the current state-of-the-art for all three segmentation tasks.

3.2.3 Railway scene understanding

Wilddash has provided data for semantic understanding of road scenes, but one important part of urban traffic has received much less attention: trams and trains. RailSem19 [ZMZ+19] has been released to mitigate this data gap and improve safety of railway applications. It is the first public dataset for semantic rail scene understanding, delivering 8500 railway scenes with rich semantic annotations. An updated version of RailSem19 with more scenes and additional annotations is currently under development, including a dedicated benchmark part with public leaderboards.

3.2.4 Redundancies in image datasets

The selection of suitable scenes for training and test datasets is time-consuming. One aspect promoted by CV-HAZOP is the selection of difficult/challenging test cases. Another important topic is the coverage of scene variability without too many redundancies. The automatic filtering of images based on image hashes [ZZ21] has provided an easy tool to reduce redundant images. Follow-up work will introduce efficient databases for image hash comparisons at scale to allow efficient use during the creation of large-scale datasets.

3.2.5 Robust Vision Challenge

Another aspect of robust validation is evaluation bias. This describes the inherent bias associated with evaluations due to using a specific benchmarking test dataset. This effect is a special concern for public benchmarks with leaderboards. Their known benchmarking datasets become the target of optimization. The scientific community uses the position in leaderboards to validate the quality of their approaches. Due to evaluation bias, there is an inherit benefit in optimizing (or basically solving) the benchmarking dataset instead of the underlying task at hand. Data-driven machine learning approaches can memorize and focus on characteristics of the small benchmark data in favor of general robustness, as only the benchmarking results will be relevant to paper reviewers. In 2017 the Robust Vision Challenge (RVC) [RVC22] was devised by computer vision benchmark operators 3.1 to reduce evaluation bias. Each computer vision task in the challenge

Table 3.1: Benchmarks that previously participated in one or more Robust Vision Challenges and their associated computer vision tasks. **Obj**ect detection, **Ste**reo, **Flo**w, **Dep**th predict.; **S**emantic, **I**nstance, **P**anoptic **S**egmentation.

Name	Challenge Tasks						
ADE20K[ZZP+17]					SS		
COCO[LMB+14]	Obj				SS	IS	PS
Cityscapes[COR+16b]					SS	IS	PS
ETH3D[SSG+17]		Ste					
HD1K[KNH+16]			Flo				
KITTI[GLU12, MG15]		Ste	Flo	Dep	SS	IS	PS
MVD[NORBK17]					SS	IS	PS
Middlebury [SS02, SCD+06, BSL+11, SHK+14]		Ste	Flo				
MPI-Sintel[BWSB12]			Flo	Dep			
OID[KRA+20]	Obj						
rabbitAI[SGB+20]				Dep			
ScanNet[DCS+17]					SS	IS	
VIPER[RVRK16]			Flo	Dep	SS	IS	PS
Wilddash[ZHM+18, ZSR+22]					SS	IS	PS

(e.g. stereo vision, semantic segmentation) requires participants to submit results to multiple benchmarks but must be generated by a single solution/network. The individual subrankings are joined by a method deemed mathematically fair in electoral science: Schulze Proportional Ranking (PR) method [Sch11]. The joined meta-ranking represents a more robust evaluation with reduced evaluation bias. In addition to the changing image content characteristics, different data specifications and policies are joined at RVC:

- Depth ranges vary between the multiple stereo vision datasets
- Displacement vectors vary in scale and orientation characteristics (optical flow)
- Semantic classes differ widely between segmentation datasets (e.g. Outdoor car traffic labels for Cityscapes and Wilddash are mixed with indoor labels of ScanNet)
- Input image dimensions and aspect ratios vary

RVC thus also helps to increase the applicability of the competing solutions by requiring a richer feature set based on the union of the individual benchmark specifications. RVC has been hosted in 2018, 2020, and 2022. The main organizer for the first RVC was Andreas Geiger, all later and upcoming RVC workshops are organized by Oliver Zendel. It is planned to continue as a biennial workshop series and new ideas for further reduction of evaluation bias will be incorporated in the future. The workshop will focus on advancing meaningful evaluations for computer vision validation to test the robustness of solutions for solving real-world tasks.

CHAPTER 4

Supplemental Materials

4.1 Analyzing Computer Vision Data - The Good, the Bad and the Ugly

Supplemental Material

This is the supplemental material for the CVPR 2017 paper *Analyzing Computer Vision Data - The Good, the Bad and the Ugly* by Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernández Domínguez

Overview:

1. Visualization of algorithm results for each of the identified hazard frames for all datasets (all thumbnail images are taken from the respective datasets): Section 1
2. Cumulative calculation of average performances for all datasets: Section 2
3. Full hazard list specialized for stereo vision: Section 3
4. Identified hazards per dataset (and corresponding frame): Section 4
5. URLs of datasets (Datasets are the same as in Table 1 of the submitted paper): Table 12

1. Performance results of algorithms for all hazard frames

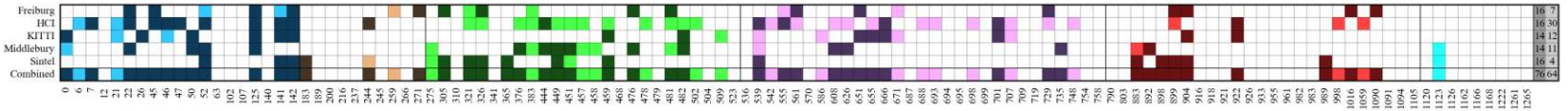

Figure 1. Overview of found/disputed entries including gaps with missing entries

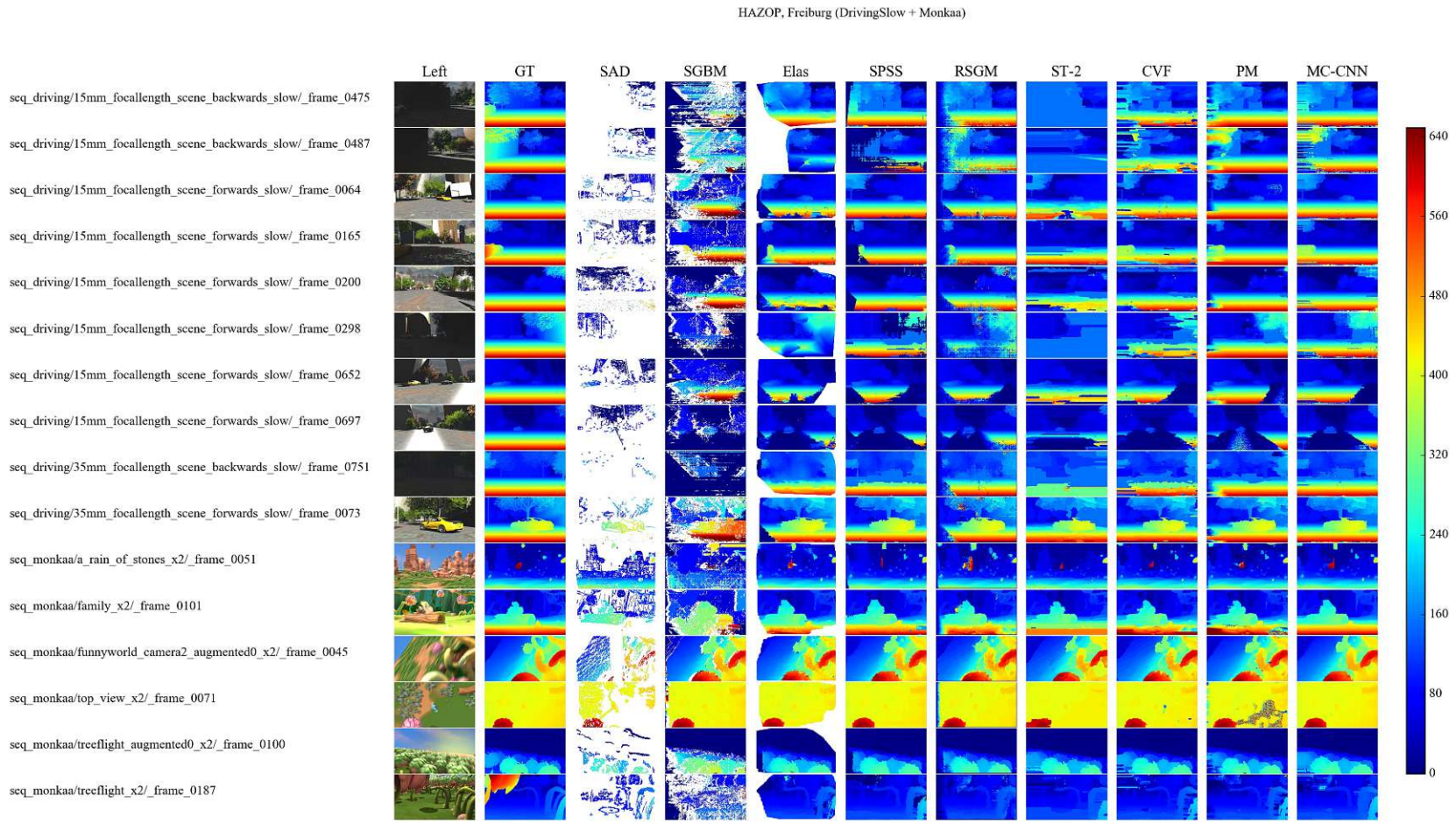

Figure 2. Disparity of each hazard frame from the Freiburg dataset

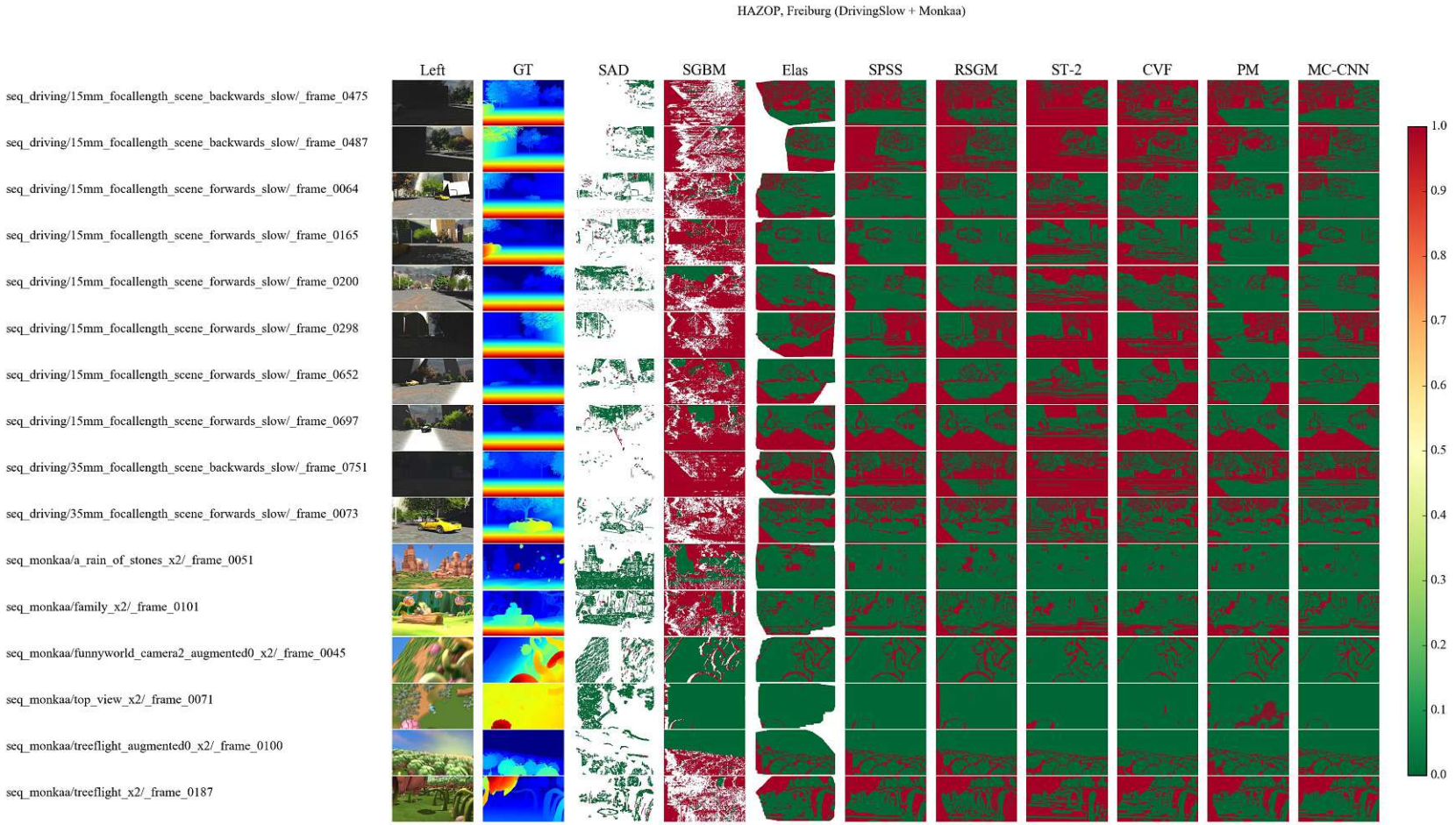

Figure 3. Pixels with an error $> 4pxl$ compared to GT; each hazard frame from the Freiburg dataset

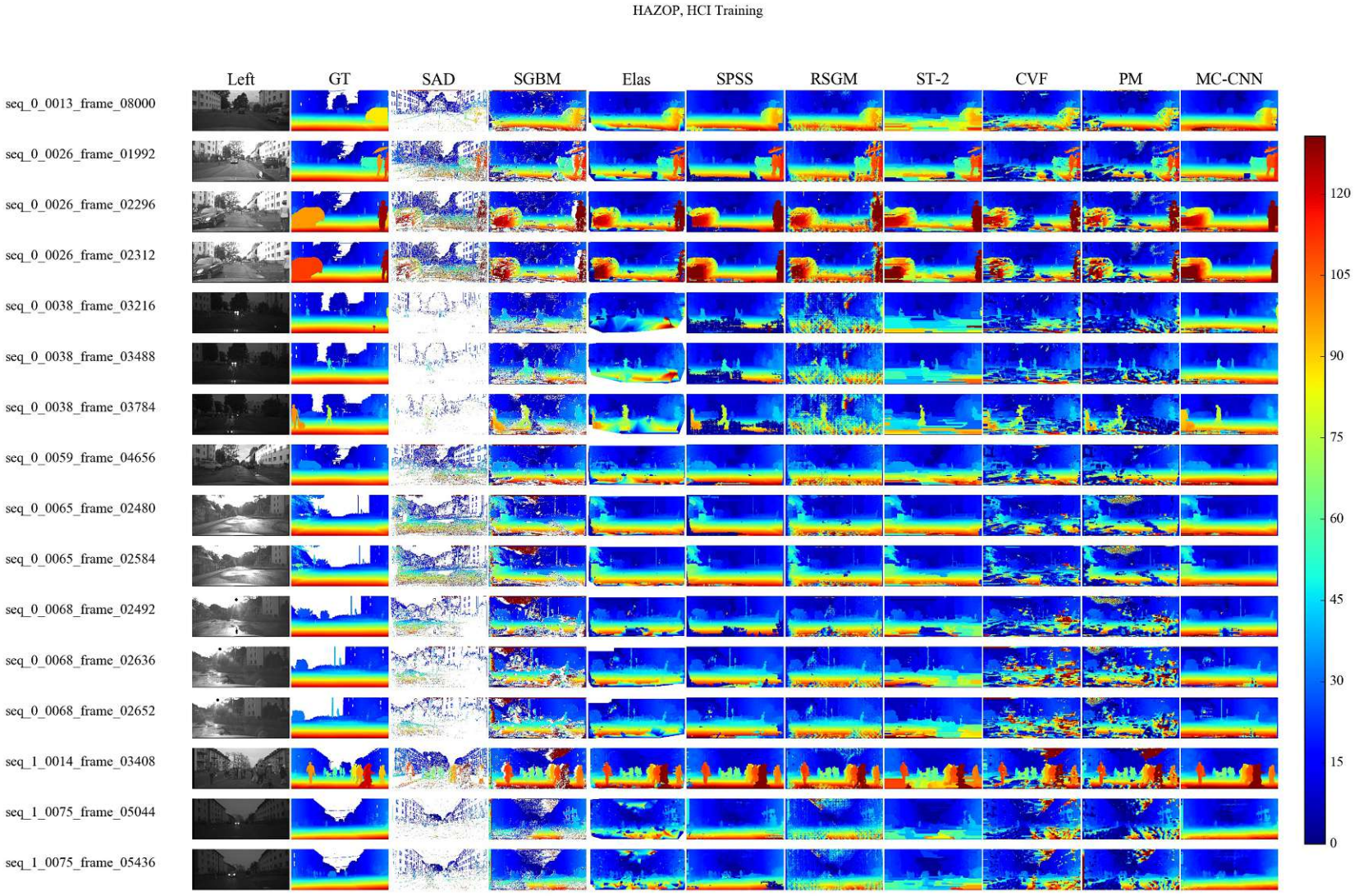

Figure 4. Disparity of each hazard frame from the HCI dataset

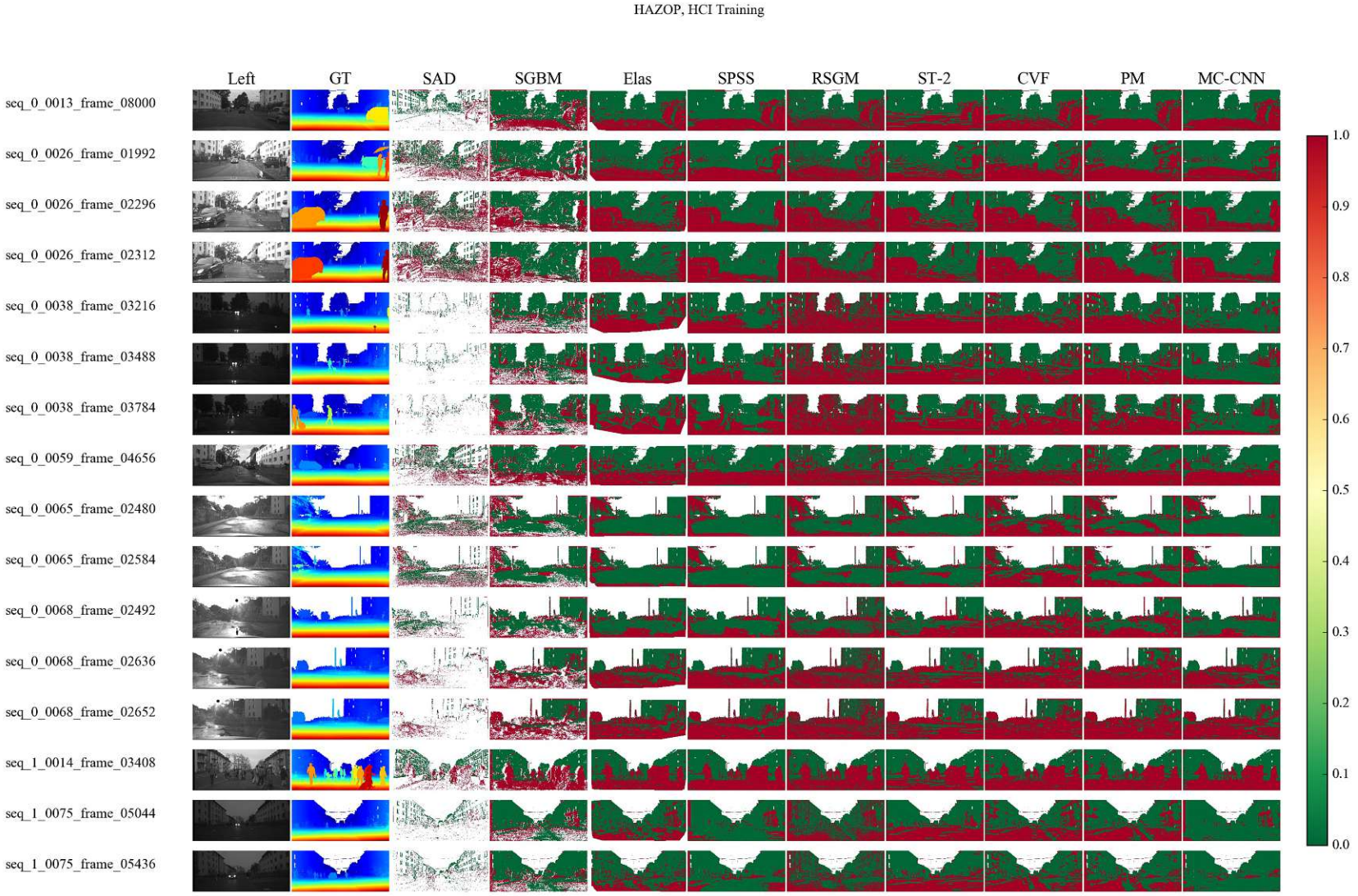

Figure 5. Pixels with an error $> 4pxl$ compared to GT; each hazard frame from the HCI dataset

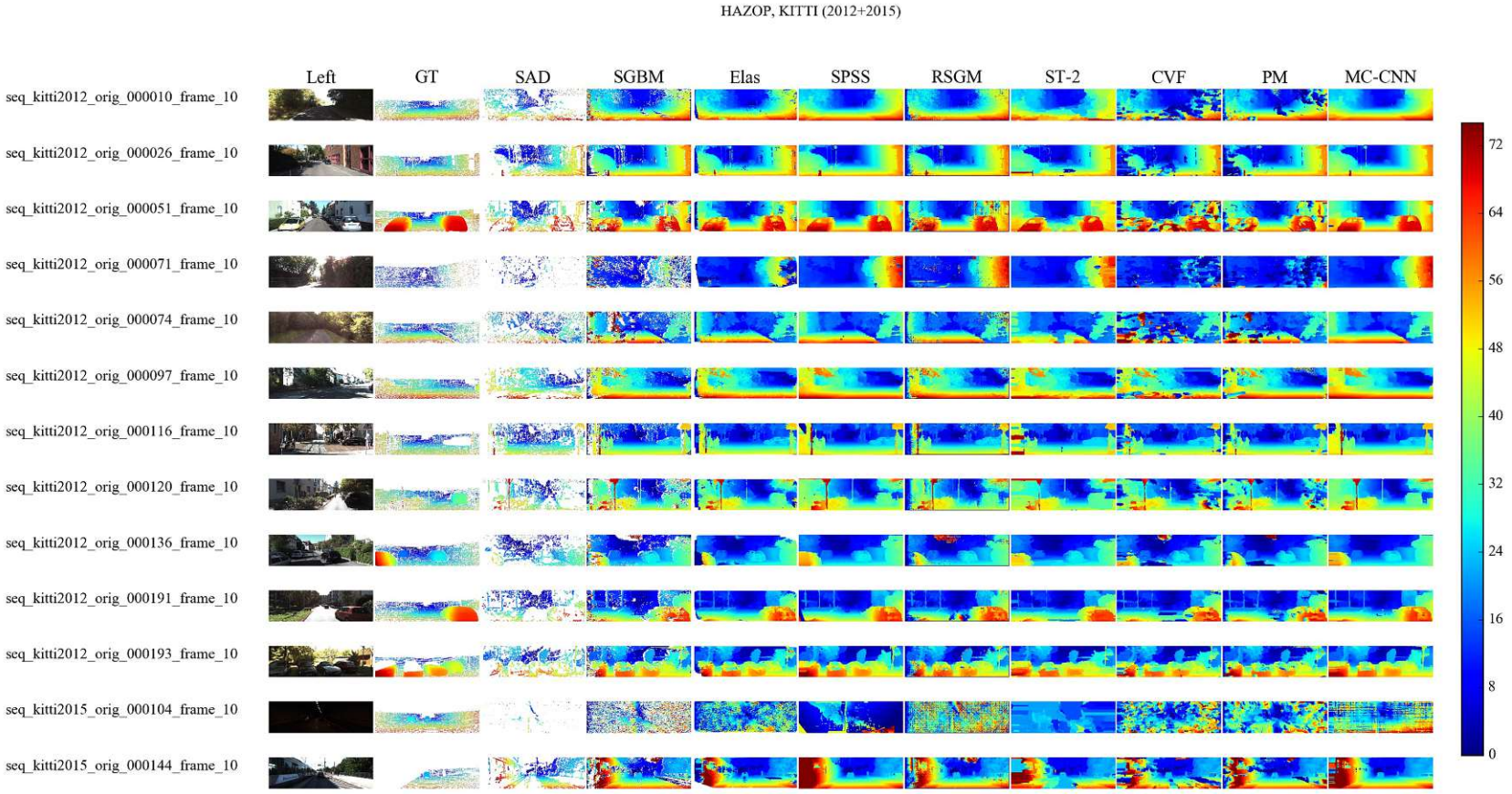

Figure 6. Disparity of each hazard frame from the KITTI datasets (2012 & 2015)

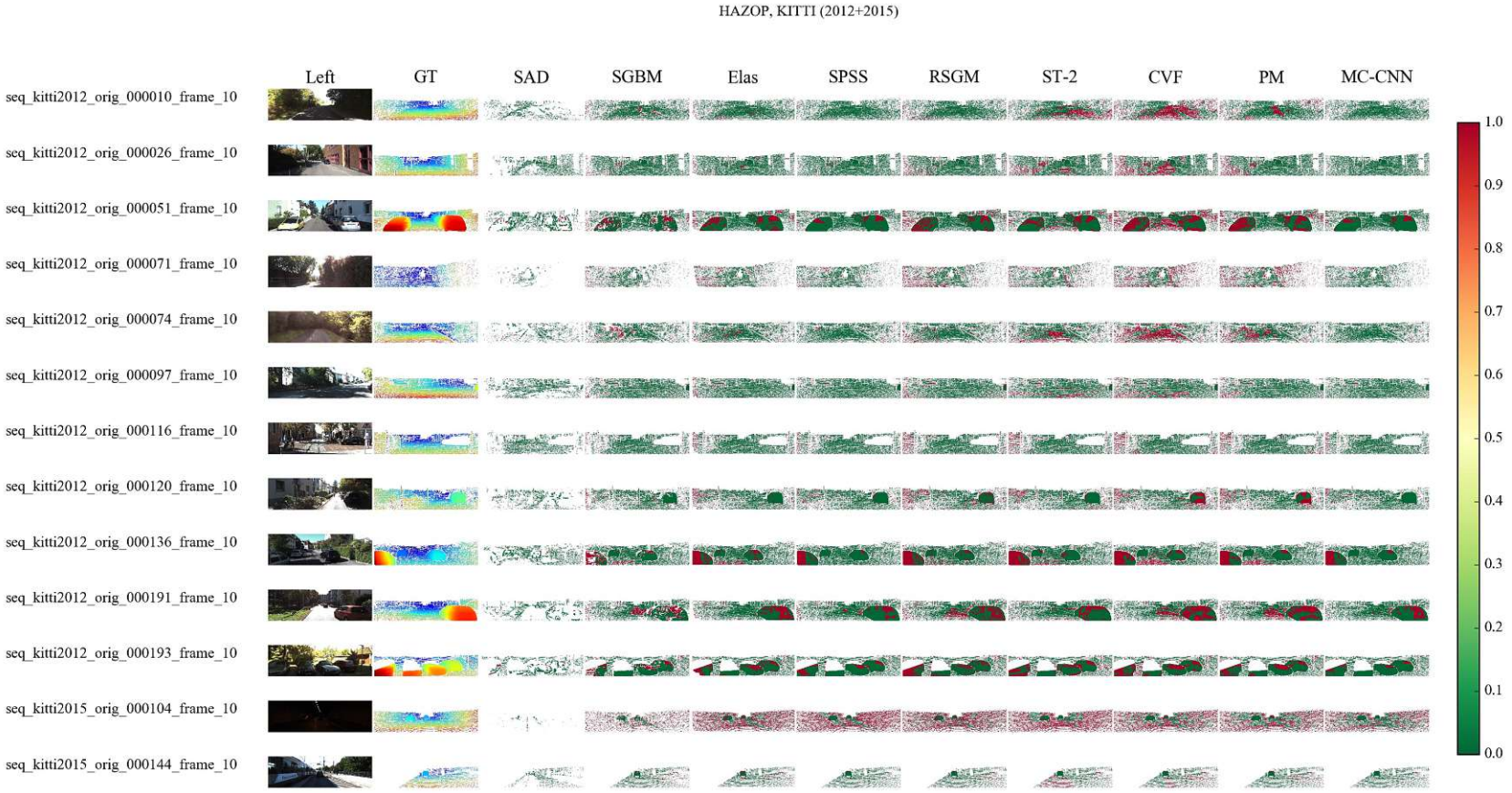

Figure 7. Pixels with an error $> 4pxl$ compared to GT; each hazard frame from the KITTI datasets (2012 & 2015)

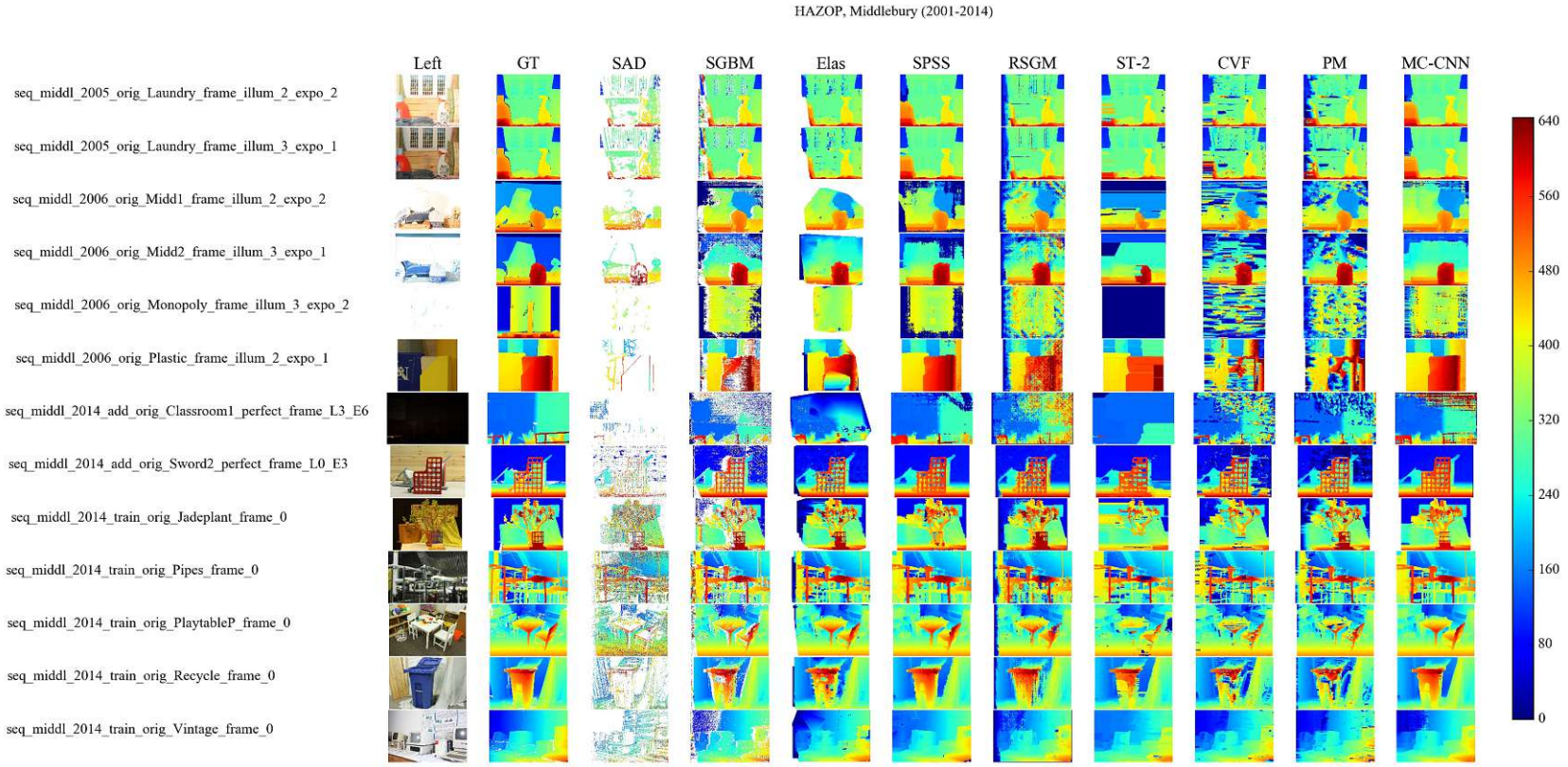

Figure 8. Disparity of each hazard frame from the Middlebury datasets

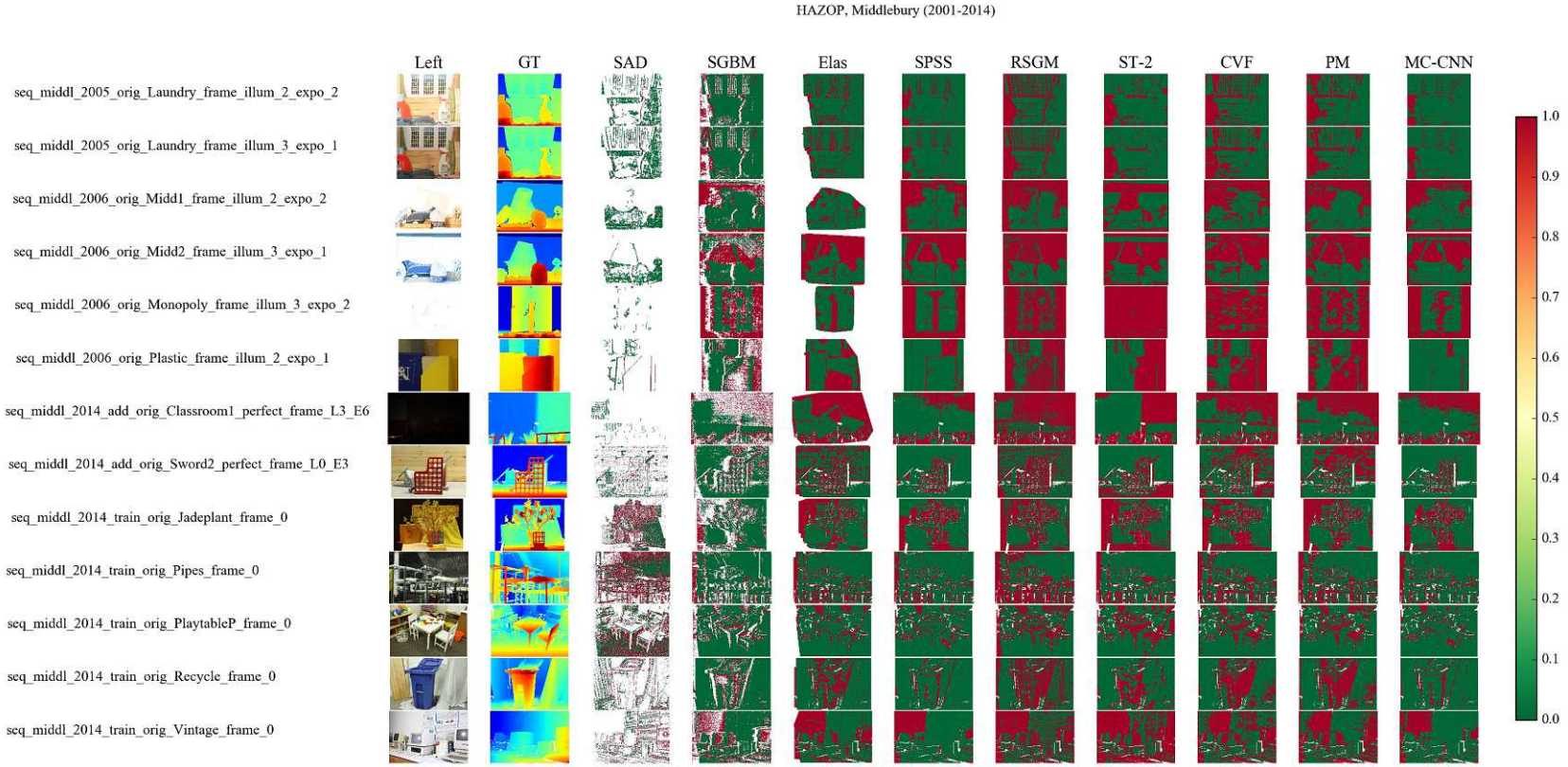

Figure 9. Pixels with an error $> 4pxl$ compared to GT; each hazard frame from the Middlebury datasets

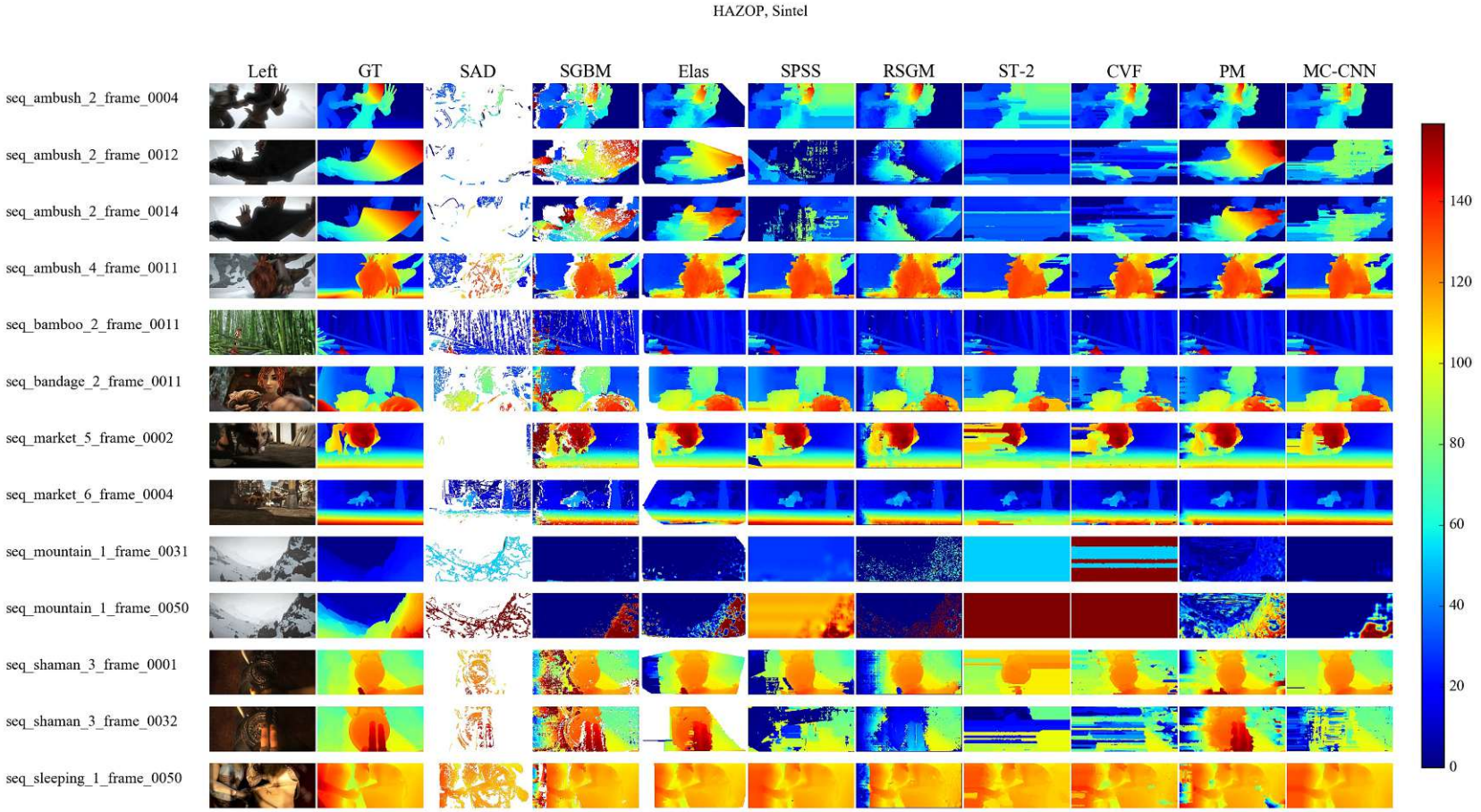

Figure 10. Disparity of each hazard frame from the Sintel dataset

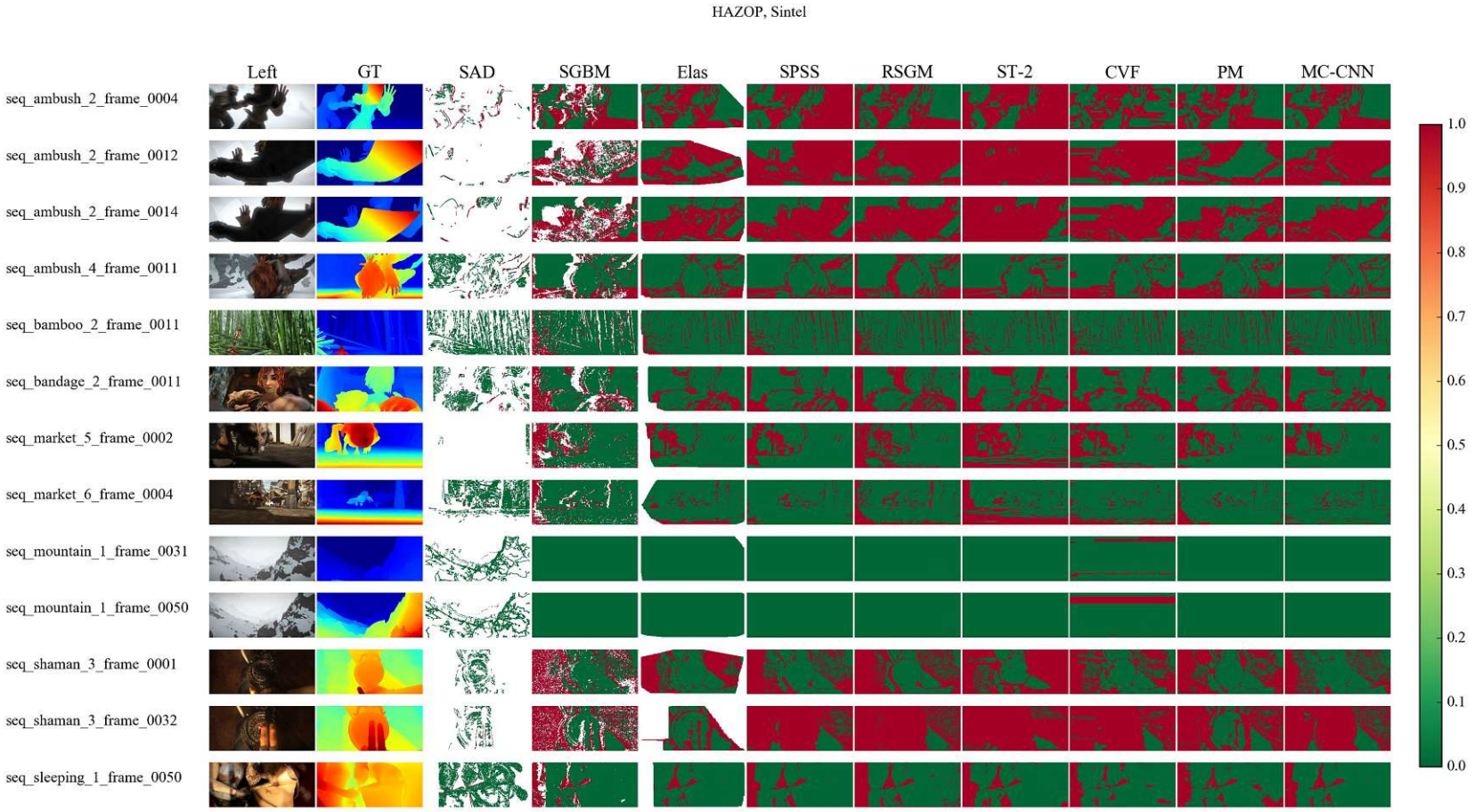

Figure 11. Pixels with an error $> 4pxl$ compared to GT; each hazard frame from the Sintel dataset

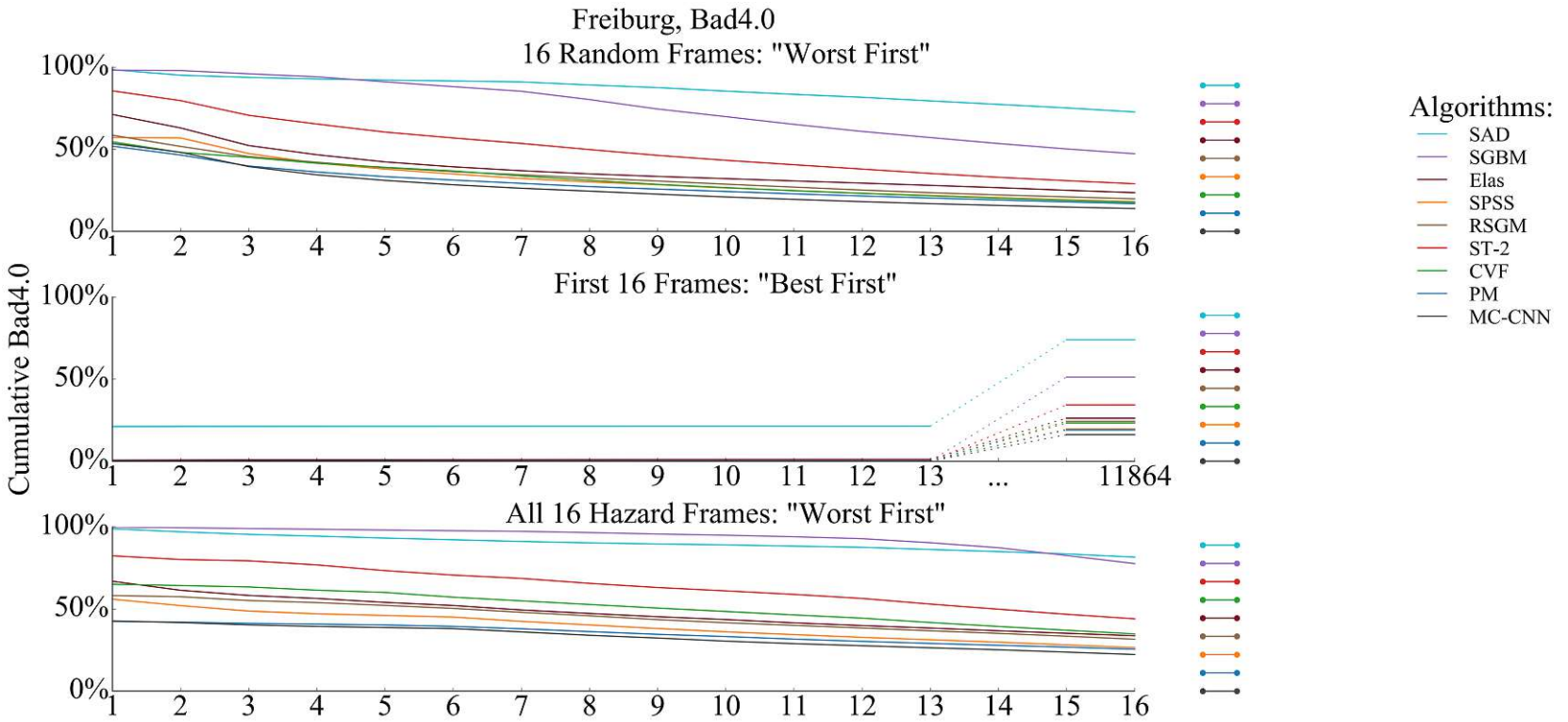

Figure 12. Comparison of cumulative average performance of 16 frames from Freiburg: Random picking, easiest frames, hazard frames (all sorted by difficulty)

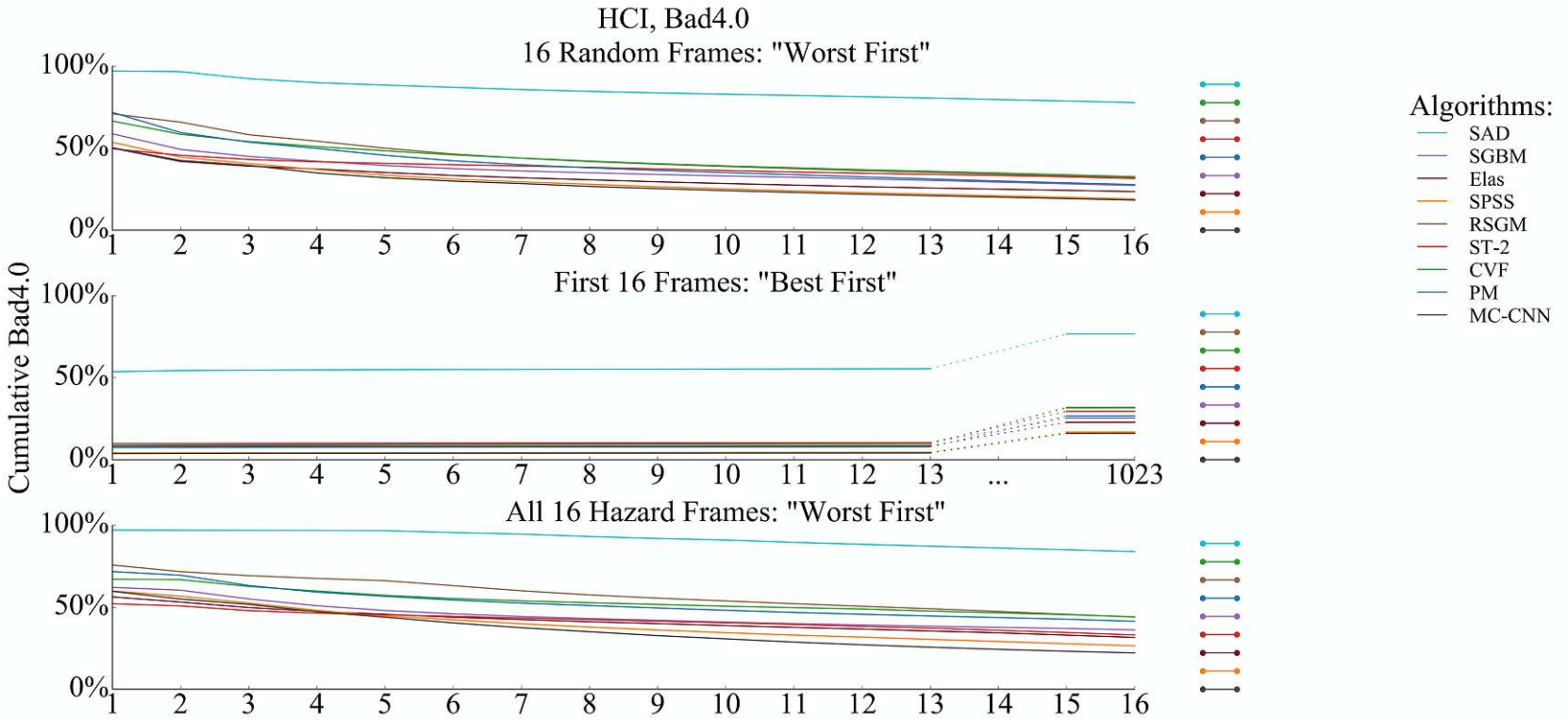

Figure 13. Comparison of cumulative average performance of 16 frames from HCI: Random picking, easiest frames, hazard frames (all sorted by difficulty)

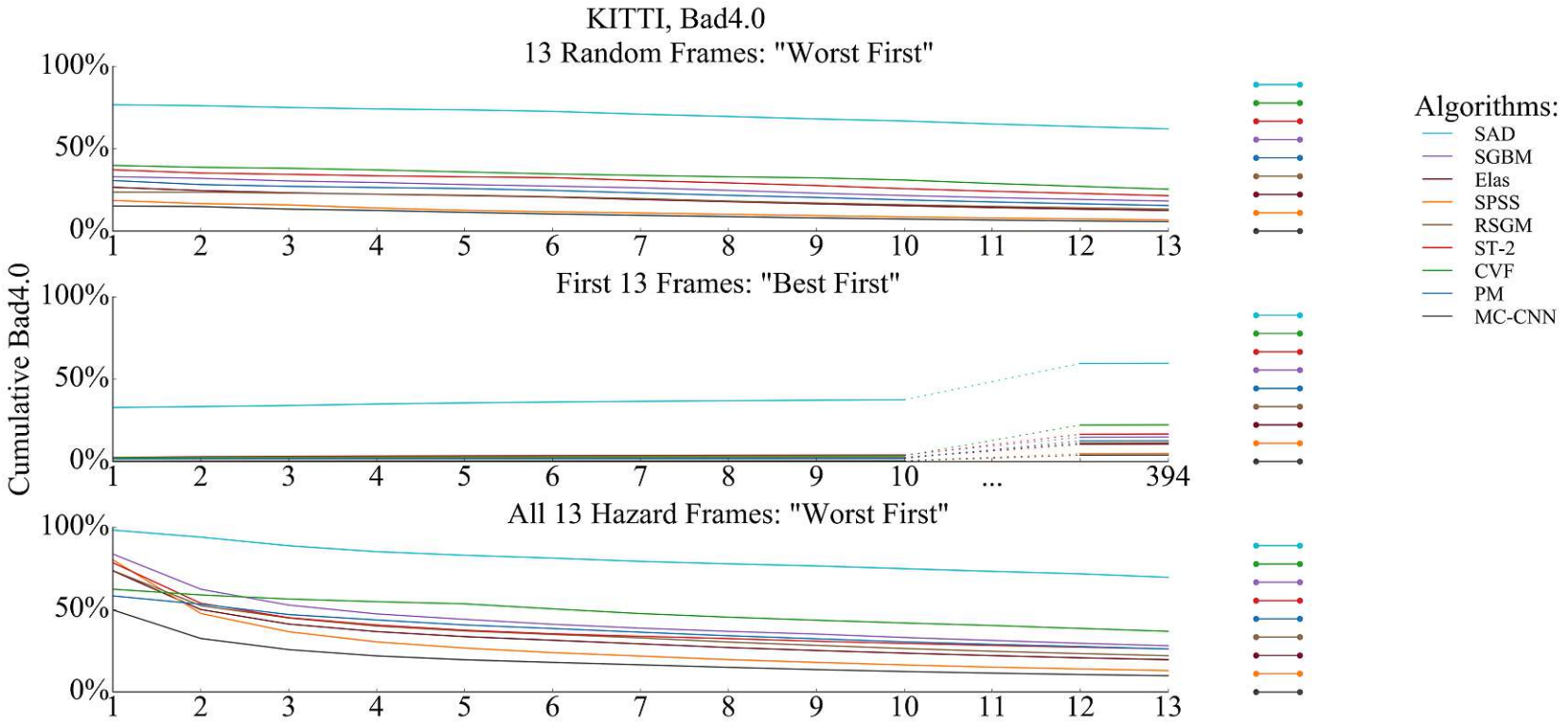

Figure 14. Comparison of cumulative average performance of 13 frames from KITTI: Random picking, easiest frames, hazard frames (all sorted by difficulty)

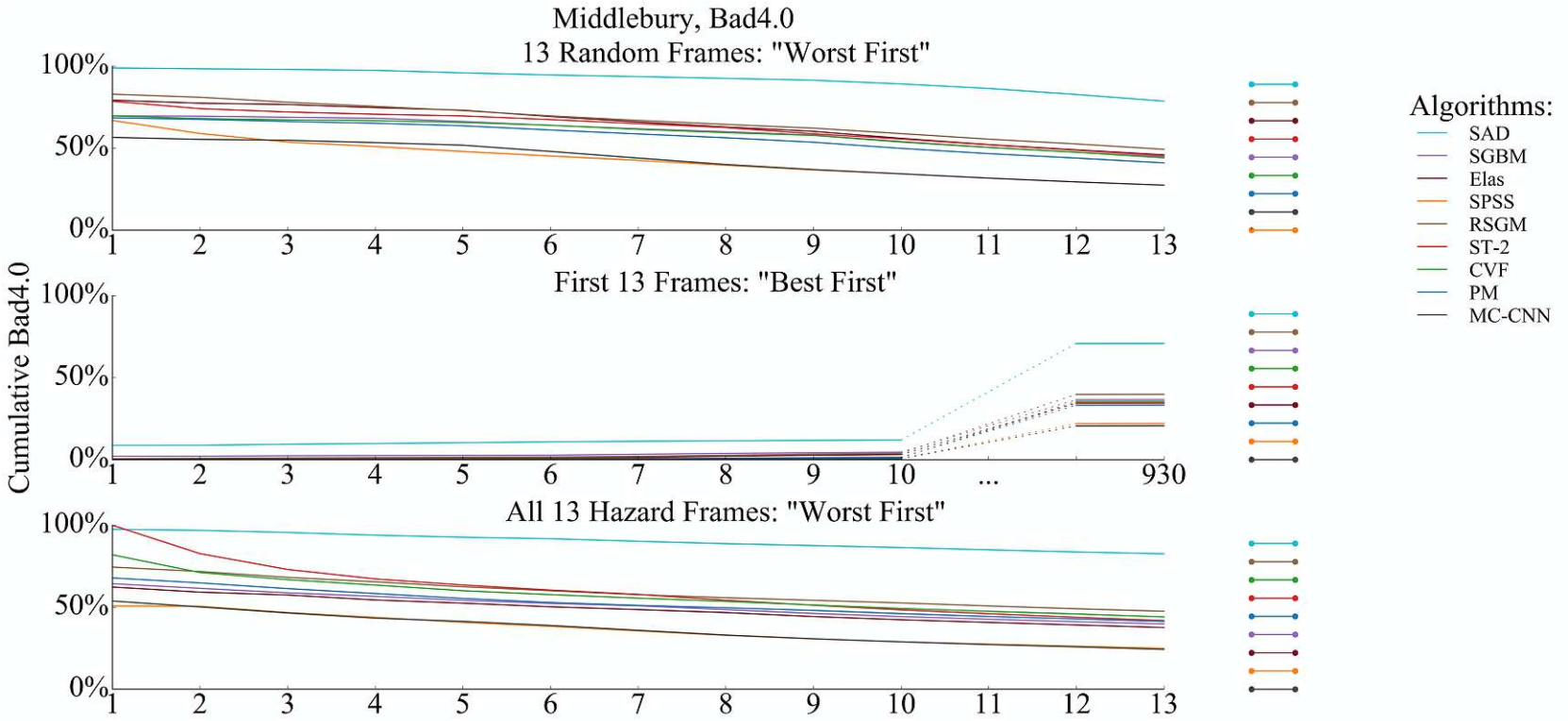

Figure 15. Comparison of cumulative average performance of 13 frames from Middlebury: Random picking, easiest frames, hazard frames (all sorted by difficulty)

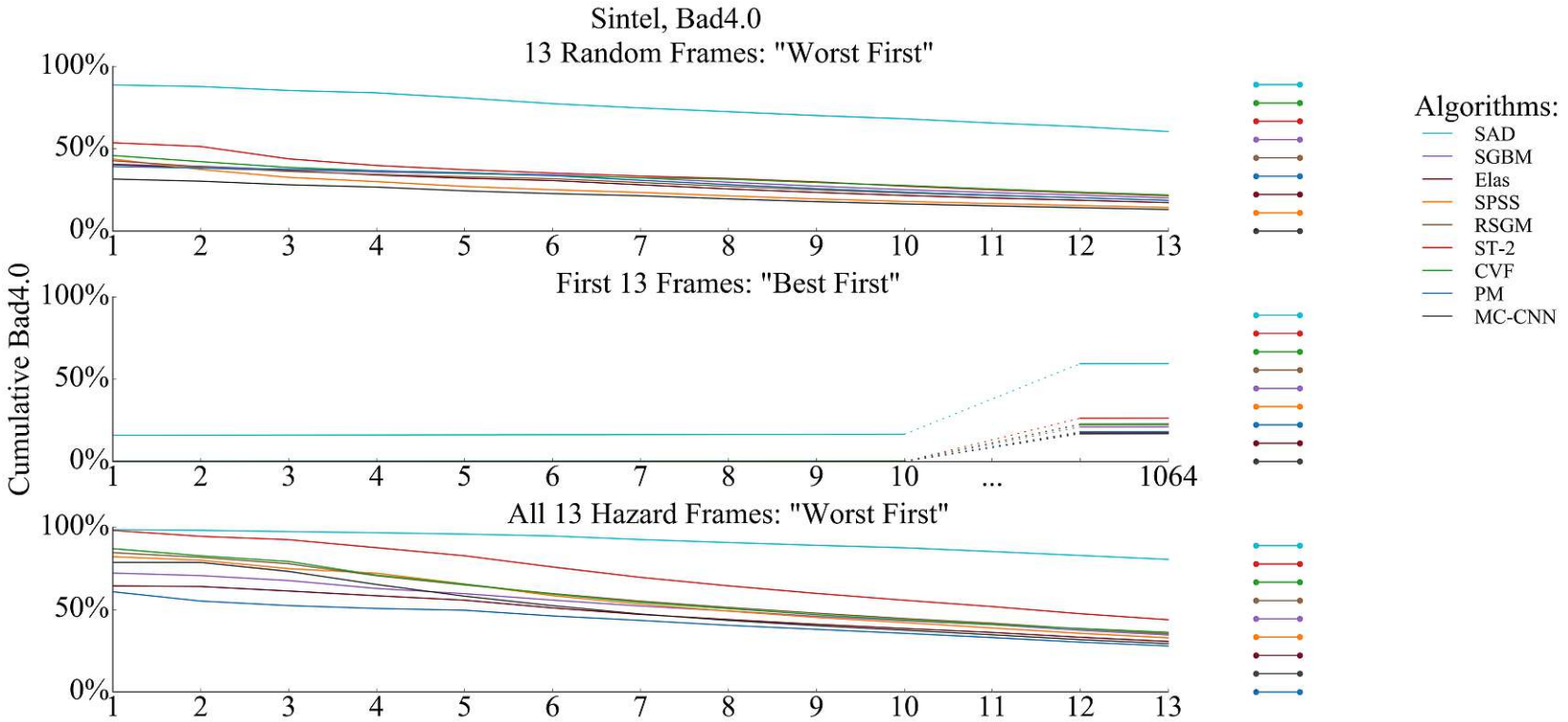

Figure 16. Comparison of cumulative average performance of 13 frames from Sintel: Random picking, easiest frames, hazard frames (all sorted by difficulty)

3. Full hazard list

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
0	L.s.	No	Number	No l.s.	No light available	Sensor will receive no light, but thermal noise or black current can cause wrong input	Highly underexposed image where black-level noise makes up most of the data; Integration test: system is self-aware that its output (or output at these areas) are not trustworthy
6	L. s.	As well as	Number	Mirrors fake additional l. s.s	L.s. can appear at locations other than where they are	Algorithm confuses position of l. s.s	L.s. as well as mirror image of the same l. s. are visible in the image. Critical example for stereo vision: L.s. and reflection are on the same epipolar line (e.g. table with candle with a large mirror directly behind it).
7	L. s.	As well as	Number	Mirrors fake additional l. s.s	Increases shadow complexity	Algorithm detects more l. s.s than exist	A l.s. and its clear reflection are near-perfect aligned on the same epipolar line
12	L. s.	Spatial periodic	Number	Several l. s.s are configured in periodic manner	Consequences depend on combined param.	Hazards depend on combined param.	There is a periodically ordered array/line of l.s. aligned on the same epipolar line for both cameras (this can occur at large distances or when aligned with the horizon-line)
21	L. s.	Less	Position	L.s. near to observer	Lighting of scene can be too strong	Over- and underexposure in same scene possible	L.s. visible in image is near to the camera and overexposed while areas surrounding the l. s. quickly get dark and under exposed (E.g. room only lit by a candle)
22	L. s.	Less	Position	L.s. near to observer	Light intensity may decrease (with increasing distance from l.s.) significantly within scene	Only parts close to l.s. sufficiently illuminated	Scene with extreme light fall-off: minority of image is well lit with a rapidly decreasing illumination around it
26	L. s.	Part of	Position	Part of l.s. is visible	L.s. at the image's edge looks different than in the middle	Overexposure (of image parts)	L.s. in image is cut apart by image border

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
45	L. s.	In front of	Position	L.s. is part of scene (in front of observer)	L.s. can be directly visible from observer	Overexposure (of image parts) - local outshining	L.s. is prominently visible in image and is surrounded by considerable overexposed areas
46	L. s.	In front of	Position	L.s. is part of scene (in front of observer)	L.s. at the imageĂŹs edge looks different than in the middle	Reflections of optics in image	Clearly visible Bokeh together with the l. s. causing it (e.g. the sun)
47	L. s.	In front of	Position	L.s. is part of scene (in front of observer)	L.s. can be directly visible from observer	Virtual rays in image	L.s. together with clearly visible streaks of light radiating in a radial fashion from L.s.
50	L. s.	Behind	Position	L.s. behind Observer	Objects illuminated with small angle between direction of light and direction of view	Small irregularities on object surfaces with same colours as surroundings may remain undetected	Scene where sun (or other strong l.s. is directly behind the observer. Relevant untextured object's structure is not reconstructed due to missing object self-shading.
52	L. s.	Behind	Position	L.s. behind Observer	Little contrasts on smooth surfaces	Reflecting areas oriented parallel to image plain may appear over exposed	Sun behind camera is casting light on a white wall causing overexposure
63	L. s.	Faster	Position	L.s. moves faster than expected	L.s. stays shorter at a place than expected	Too weak light	L.s. visible in image with a long elongated thin shape (e.g. neon tube) creating an unusually prolonged overexposed area
102	L. s.	More	Texture	L. s. has too much texture	The l. s. produces a texture of its own by projecting a textured light beam (virtual texture)	Texture of emitted light is confused with texture on object. This creates false positive detections.	L.s. projects a texture onto a surface while a very similar texture is already present next to it as part of another object's surface texture, both textures are aligned on the same epipolar lines

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
107	L. s.	As well as	Texture	L. s. projects combination of two textures, one expected, the other unexpected	Blending of lightings, complex illumination and shadowing	Small changes in l.s. configuration may cause large differences in responses of CV algorithm (e.g. Moire)	L.s. projects a thin structured pattern onto a surface that produces two distinctly different Moire patterns in the left/right camera
125	L. s.	More	Intensity	L.s. is too strong	Too much light in scene	Overexposure of lit objects	Directly lit object is overexposed in an otherwise correctly exposed scene/image
140	L. s.	More	Beam	Large beam angle, even omni-dir. emission of light	All objects will be lit	Reflections in all shiny surfaces possible	Very bright scene without overexposure but very little contrast due to approximating an ambient lighting situation with nearly no shadows (self-shading neither)
141	L. s.	Less	Beam	Focused beam	Only fractions of objects will be lit	Large parts of scene may be dark	Headlight situation with only a small part of the scene sufficiently being lit. Large parts are underexposed.
142	L. s.	Less	Beam properties	Focused beam	Only fractions of objects will be lit	Unsmooth illumination of surfaces	Scene where a prominent object is only half lit by the scene's l.s. while a large portion remains severely underexposed
183	Medium	Less	Transparency	Medium is optically thicker than expected	Less light can pass through	Less contrast than expected could result in mismatches	Fog / haze in image reduces visibility depending on distance from observer
189	Medium	As well as	Transparency	Two different media have a different optical thickness	Refraction occurs: changes the path of light from the object to the observer	The object appears to be displaced	There is a large part of the scenery clearly visible within a different medium than in directly in front of the observer (e.g. view clean/clear water with lots of details visible beneath the water surface)
200	Medium	As well as	Spectrum	Medium has similar colour as nearby l.s./object	Low contrast	Objects and medium become indistinguishable	Scenery contains a medium (air, water) with comparable colour and particles/textures as the objects in the scene
216	Medium	Spatial periodic	Texture	Texture of medium is periodic (periodic density fluctuations)	Medium projects periodic texture onto surfaces	Confusion with object texture by CV alg. possible	A periodic appearing density fluctuation creates a periodic pattern on a visible surface aligned with the stereo system's epipolar geometry

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
237	Medium	No	Particles	No particles in the medium	No particles in the medium which scatter transmitting light	If particles are needed to e.g. visualize flow dynamics, this will be hampered	The border between two media is very clean and the medium is clean as well thus preventing the detection of the medium border itself
244	Medium	More	Particles	Particles are large(r than expected)	Particles appear as distinct objects	Particles are misinterpreted as objects	Large hailstones, snowflakes or raindrops look like parts of the actual scene/objects in the scene thus creating faulty matches
245	Medium	More	Particles	Particle size is bigger than the light's wavelength	Geometric Scattering	See Less Trans or More Texture	Cloud of visible particles (e.g. pollen, small leaves) in the air are obscuring the scene
259	Medium	Where else	Particles	Particles fill up different parts of the scene with different density	Different areas of scene exhibit different visual effects	Different recognition quality throughout an image	Scene is split into two roughly equally big parts: one without particles and another with considerable amount of particles (e.g. a view with a roof covering a area where no snow/rain is falling and an outside part full of rain/snow)
266	Medium	Close	Particles	Particles very close to Observer	Single particles may cover larger scene fractions	Single particles are confused with real scene objects	A single particle that is close to the observer looks very similar to an object in the scene while both are aligned on the same epipolar lines
271	Medium	Faster	Particles	Particles move faster than expected	Motion blur of particles	Blurred particles obfuscate (parts of) scene	Scene contains particles moving fast enough to have a noticeable motion blur
275	Object	No	Position	Pos. cannot be defined/ detected	An object's "central" point cannot be defined	One object is reported as several	Large, diffuse or highly structured or flexible objects like clouds, fungus mycelium, or table-cloth is broken into many objects small enough so that noise/speckle filtering might remove them
305	Object	Faster	Position	Obj. moves faster than expected	Obj. stays shorter at a place than expected	Transversal motion blur	Object is moving from left to right fast enough to have a noticeable motion blur

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
310	Object	More	Size	Obj. is larger than expected	Obj. has a size more similar to other objects than its own characteristic size	Object is confused with some other object	Two very similar objects but of different scaling are present in the scene. The two objects are positioned on the same epipolar lines and their projected size is the same (due to perspective effects)
321	Object	Part of	Size	Only Part of Obj. area is visible	Obj. is either partially occluded	Object is not correctly recognized	Larger parts of an object are occluded (left view vs. right view) so that the remaining parts might get rejected as noise/speckles
326	Object	Part of	Size	One of the Object extents is missing	Degenerated configuration of object surface	CV alg. fails because of a degenerated case	A large but very thin object is positioned in such a way that exactly one of the two cameras sees only the thin edge of it without much surface.
341	Object	Faster	Size	Obj. size changes faster than expected	Obj. shrinks/increases/pulses remarkably during exposure	Radial motion blur	Scene contains an expanding/shrinking object that has a noticeable radial motion blur (not caused by ego-motion!)
365	Object	Faster	Orientation	Orient. changes faster than expected	Obj. rotates remarkably during exposure	Rotational motion blur	Scene contains a rotating object that possesses noticeable rotational motion blur (not caused by ego-motion!)
376	Object	Less	Complexity	Object is less complex than expected	Object lacks natural features	Insufficient amount of natural features leads to faulty/no results in 3D reconstruction or self-localisation	Simple non-planar object without texture or self-shading (e.g. grey opaque sphere)
383	Object	Other than	Complexity	Object has a complete different complexity (shape) than expected	Parts of object are identical	Mismatch of object parts in stereo lead to wrong depth or shape recognition	Locally simple repeating parts of an object that is otherwise complex (e.g. house facade with a regular grid of windows)

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
444	Object	No	Texture	Object has no texture	Object is either monochrome, or is highly reflective or transparent	Texture-based CV alg. will not work	Large parts of the image are completely textureless
449	Object	Less	Texture	Object has less texture than expected	Texture is no significant identification property	Texture-based CV alg. is hampered	Two objects at the same image height (on epipolar lines) have very little texture thus allowing a mismatch
451	Object	As well as	Texture	Object has a mixture of periodic and aperiodic texture	Semi-periodic texture: periodic at coarse LOD, but differences in detail	Object not correctly recognized	Texture of a large object is periodically repeating on a coarse level while small local variations exist
457	Object	Spatial periodic	Texture	Obj.texture is periodic	Same appearance of texture on different parts of objects	Object parts are confused	Texture on epipolar line is highly repetitive (coarse and detail level) and mostly normal to the screen (nearly no perspective distortion)
458	Object	Spatial periodic	Texture	Obj.texture is periodic	If the period undercuts half the resolution of the observer spatial aliasing occurs	Stereo alg.s compute wrong depth maps due to mismatch of texture cells	Fine-structured texture undercuts spatial resolution of the sensor. Aliasing artifacts create a different moire patterns for the two images
459	Object	Spatial aperiodic	Texture	Obj. texture is aperiodic	Texture does not precisely repeat, but variations are irrelevant	Irregular (stochastic) mismatches in stereo images	A large area is loosely periodically tiled (w.r.t. epipolar lines) but the tiling is not perfect (e.g. floor tiling with some variations)
468	Object	Remote	Texture	Two very different textures are touching on the same object	Borders between the textures create strong contrast there	Object recognition hampered, if border between textures creates new visual artifacts	Two prominent textures on the same object are creating a very distinct border region where they collide. This border region is only visible by one camera but occluded for the other

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
476	Object	No	Reflectance	Obj. has no reflectance	No light reflected	Object confused with shadow	Well-lit scene contains a very dark/black object that appears to have has neither texture nor shading due to its low albedo
478	Object	More	Reflectance	Obj. has much Refl. (more than expected)	Shiny surface - mirror	Object not recognized	Object has strongly reflecting material that creates an arbitrary mirror-image as the object's texture
479	Object	More	Reflectance	Obj. has much Refl. (more than expected)	Overexposure of the observer	Reflected objects taken for real	Object has strongly reflecting material that mirrors larger parts found on the same epipolar line
481	Object	As well as	Reflectance	Obj. has both shiny and dull surface	Diffuse reflection with highlight/glare	Object recognition distorted by glares	Object has a large glare spot on its surface that obscures different areas in the left/right image; this happens when the glare inducing l.s. is positioned near to object and observer
482	Object	As well as	Reflectance	Obj. has both shiny and dull surface	Diffuse reflection with highlight/glare	Local overexposure due to glares	Object has a large glare spot on its surface that obscures same areas in the left/right image
502	Object	More	Transparency	Obj. is highly transparent	Transparent object	Object not recognized	Highly transparent empty object covers large parts of the scene, the scenery behind the object is clearly visible
504	Object	More	Transparency	Obj. is more transparent than expected	Transparent object	Objects within it not correctly recognized due to distortions, e.g. through glass	Highly transparent object encompassing a second opaque object that gets distorted due to the transparent object's shape
509	Object	As well as	Transparency	Obj. is both more and less transp. than expected	Obj. consists of parts with high and low transparency	Object itself and objects behind it are merged	Scene contains a large object with a mixture of high transparency and low transparency. The object and the scenery behind it are close to both cameras so that occlusions occur
523	Object	In front of	Transparency	Objects in front of a transparent object are visible together with the other object	Objects in front of a transparent object are visible together with the other object	Objects are confused	Two transparent objects are entangled in such a way that both allow the view on the other object. They are arranged on epipolar lines so that different parts might get faulty correlations

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
536	Objects	No	Number	No objects	Scene with no objects (only l. s.s and media)	Non-existing objects might erroneously be reported by CV alg.	Scene without visible objects of any kind; only homogeneous medium and pixel noise is visible
539	Objects	More	Number	More objs. than expected	Scene is more complex than expected	False negatives: objects are missed	Scenery made up of many individually different objects at different distances that clutter the scene and create a highly variating disparity range
542	Objects	More	Number	More objs. than expected	Scene is more complex than expected	An object is covered such that uncovered parts are interpreted as belonging to different objects	Two objects occlude different parts of two other objects. The occluded parts are the exact/near copies of the vice-versa occluded part -> the first object occludes something that the second occluder reveals (and vice-versa)
555	Objects	Spatial periodic	Number	Object arrangement is periodical	Observer resolution and windowing have to be appropriate to capture a characteristic arrangement	If resolution of field of view are not appropriate detection based on characteristic arrangements is corrupted	Highly periodic placement of identical objects along the epipolar line creates repeating structures which lead to potential mismatches
561	Objects	In front of	Number	A number of obj. is in front of each other (in respect to the observer)	They cover each other	They are indistinguishable	Identical objects are arranged in such a way that one of the objects completely covers the other object in one of the images. Thus the covering object can create faulty matches with the covered object
570	Objects	More	Positions	More discrete relative Pos.s of Obj.s than expected	Object positions are quantized at a greater resolution than expected	CV alg. expects discrete positions but gets positions in between them confused	Subpixel-Accuracy can be tested by special test data sets (e.g. images with exactly 0.5 or 0.3333 pxl of disparity etc.)

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
586	Objects	Spatial periodic	Positions	Objects are located regularly (different kind)	Different kind of objects appear in a geometrically regular pattern	Only regularity detected, but not the individual objects	Similar objects (but not identical) are arranged in a highly periodic fashion on the epipolar line
608	Objects	More	Occlusion	More objects occlude each other than expected	Less details of objects are visible	Detection quality is decreased by less information of needed Objects	Some objects are positioned at two distinct distances. The frontal objects create considerable occlusions that might prevent the correlation of the backside objects
626	Objects	Spatial aperiodic	Occlusion	Occl. creates a chaotic /unordered pattern	Occlusions are chaotic	CV alg. is not handling occlusions correctly	Scene is dominated by an aperiodically perforated object near to the observer thus occluding many parts of the scene behind the object
651	Objects	More	Shadowing	More shadowing than expected	Large parts of scene in shadow	Underexposure: objects in shadow not detected	Large parts of a well lit scene are underexposed due to large shadows cast by objects not seen in the scene.
655	Objects	Less	Shadowing	Less shadowing than expected	More parts of scene in light than usual	Overexposure: similar to No	Largely open space with little shadows creating an overexposed lighting condition
666	Objects	Where else	Shadowing	Reflecting obj. is within shadow	Reflecting object is less recognizable than if illuminated	Reflecting objects in shadow remain undetected	Scene contains a highly reflective object in a shadowed area which reflects a well-lit object into both cameras. Both objects are relatively near to the two cameras so that the reflection of the object appears at different parts of the mirroring object
671	Objects	Spatial periodic	Shadowing	Spatial periodic shadows, there is some order/rule as to what parts of an object are shaded	Regular shadows creates a pattern	CV alg. confuses shadow pattern with object	Highly periodic shadows along epipolar line creates repeating structures which lead to potential mismatches

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
687	Objects	More	Reflectance	There are more reflections between objects than expected	Creates multiple views within the scene	CV alg. might confuse reflectance with reality and infer wrong position/relation data	Both object and its clear reflection are visible on the same epipolar line (and same distance) in both left+right image. The mirrored object is symmetric (mirror image looks like the original object) which can lead to a faulty correlation.
688	Objects	More	Reflectance	There are more reflections between objects than expected	Can create multiple visible instances of the same object	CV alg. detects more objects than there are in the scene	Both object and its clear reflection are visible on the same epipolar line (and same distance) only in one of the two images (the other image shows only the object itself). The mirrored object is symmetric (mirror image looks like the original object) which can lead to a faulty correlation.
693	Objects	Part of	Reflectance	A reflection on an object is partially visible	A highlight in an object is partially covered by another	Overblending - partial hampering of correct situation recognition	A prominent glare spot is only visible in one of the two images
694	Objects	Reverse	Reflectance	Observer sees itself in a reflection instead of expected object	Own body/ observer itself visible as an object	CV alg. confuses observer/its own body with other objects	Scene contains a clear reflection of observer (e.g. camera head, measurement vehicle) that is epipolar aligned with objects/parts that look like parts of the observer thus leading to a potential mismatch
695	Objects	Reverse	Reflectance	Refl. are reversed	Object reflections appear reversed to expected, e.g. upside down, or laterally inverted	CV alg. confused	Scene contains a large concave mirror that shows an clean upside-down copy of parts of the scenery
698	Objects	Where else	Reflectance	Refl. Obj. is Transp.	On Obj.s. surface, reflected and seen through objects merge	Misinterpretation of reflecting object and its associated images	Objects surface shows a blend/mixture of clear reflectance as well as transparently parts behind the image

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
699	Objects	Spatial periodic	Reflectance	Refl. creates an ordered pattern	Ordered reflectance creates a pattern	CV alg. confuses reflectance pattern with object or textures	Highly periodic clear reflection of the same object along epipolar line creates repeating structures which lead to potential mismatches
701	Objects	Spatial aperiodic	Reflectance	Refl. creates a chaotic /unordered pattern	Refl. is chaotic/irregular	CV alg. confused by irregular reflectance -> misdetections	Large parts of the image show an irregular specular reflection (mirror-like but with lots of distortions; not a diffuse reflectance)
707	Objects	Close	Reflectance	Refl. Obj. is closer to Observer than expected	Reflections are larger and/or brighter than expected	Overexposure: reflection too bright	A large prominent glare spot is created by a l. s. right next to the observer (but not directly visible on the images)
709	Objects	Remote	Reflectance	Refl. Obj. is more remote from reflected object than expected	Association between real obj. and mirror image lost	False positive. mirror image reported as real object	Both object and its clear reflection are visible on the same epipolar line in both images but are positioned at different distances. The mirrored object is symmetric (mirror image looks like the original object) which can lead to a faulty correlation.
719	Objects	Behind	Reflectance	Refl. obj. behind observer	If also a reflecting object in front of observer, infinite reflections can occur	CV alg. confused	Observer is placed between two large parallel mirror facing each other so that "infinite" number of reflections occur
729	Objects	Part of	Transparency	Parts of an object are transparent and allow a part of another object to be seen	Complex mixture of multiple objects visible through projection although the objects are not intertwined	Misdetection of objects as appearances are changed	Object has transparent parts that show a different object while other parts remain opaque
735	Objects	Spatial periodic	Transparency	Transp. creates an ordered pattern	Regular transparencies in scene	CV alg. confuses transparency pattern with object	Highly periodic placement of windows/holes/clearings along epipolar line creates repeating view of a uniform background which lead to potential mismatches

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
748	Objects	In front of	Transparency	Transp. obj. in front of another Transp. obj.	Transparency effects accumulate	Objects not correctly separated	Two transparent objects are positioned behind each other so that the scenery behind the last object is still clearly visible (e.g. looking through two windows in series)
754	Objects	More	Wave	More interferences between objects than expected	Accumulation of optical effects such as halos, rainbows, auras etc.	CV alg. confused - misinterpretation of visual effects	Scene contains a prominent rainbow effect (mist/haze with a view dependent colour band). This normally only occures with a strong l.s. in behind the observer
758	Objects	Spatial periodic	Wave	Spatial periodic variation of Wave effects (of some objects)	Interferences occur regularly in scene	Confusion of objects causing interference effects	Scene contains pronounced refraction rings (e.g. oil slick)
790	Obs. Opto.	Close	Number	All observers are close to each other (short baseline)	Short baseline makes triangulation results less accurate since the displacement of corresponding images pts. is smaller	Camera pose estimation fails or is inaccurate	Easy to produce by supplying the same images for left/right
803	Obs. Opto.	More	Field of View	Observer uses a bigger FOV than expected	Focal length smaller than expected	More distant entities not detected	Scene has a wide FOV (>135deg)
883	Obs. Opto.	Part of	Viewing position	VPos is Part of scene (within scene)	Sensor too close to scene - scene partially defocused	Defocused objects not correctly recognized	In a scene with considerable depth of field: slightly near objects visible by both cameras are out of focus (near plane)
892	Obs. Opto.	Spatial aperiodic	Viewing position	Observer position is not constrained (perhaps within a given range)	Additional uncertainties due to arbitrary position of observer	Additional uncertainties introduce additional uncertainties for the position estimation of objects	Relative position between cameras slightly changed compared to their initial positions/orientations; Extrinsic calibration is thus slightly off

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
898	Obs. Opto.	Remote	Viewing position	VPos is more remote from scene than expected	Object distance is bigger than expected (out of focus)	Relevant scene details not recognized	In a scene with considerable depth of field: slightly distant objects visible by both cameras are out of focus (far plane)
899	Obs. Opto.	Remote	Viewing position	VPos is more remote from scene than expected	Object have less details than expected	Objects distances estimated less accurate	Scenery is in focus but all parts are far away (only small disparities)
904	Obs. Opto.	Faster	Viewing position	Observer moves faster than expected	Motion blur more likely with longer exposures	Blurred objects misdetected	Image has parts with clearly visible motion blur
916	Obs. Opto.	Part of	Transparency	Part of optics are less transparent than expected (e.g. dirt on lens)	Defocused areas	Misinterpretation due to thick dust irregularly distributed on lens surfaces	One camera lenses contain dust/dried mud that creates a partially defocused area in the image
918	Obs. Opto.	Part of	Transparency	Observer block part of the image	Parts of the image are black	(Partially) Blocked Objects are not detected	Lens body/lens hood is prolonged and its corners are thus blocking the view
921	Obs. Opto.	Other than	Transparency	The transparency of sensor optics is completely different from expected, e.g. due to broken lenses	The scene looks completely differently than expected, e.g. parts of it are multiplied	Strong confusion of CV alg. if fault not detected	Lens is broken cleanly through parts of the center region, apart form the crack the remaining image is clear and sharp
922	Obs. Opto.	Where else	Transparency	Lens body is not completely light proof, light can reach sensor from the side of the body	Flare effects	Overexposure of parts of the scene	Image has pronounces flare effect visible without the emanating l. s. associated with it

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
926	Obs. Opto.	Spatial aperiodic	Transparency	The transparency of the lenses changes in a spatially irregularly manner,	The intensity of the image is irregularly reduced	In particular, dense stereo vision is significantly hampered, since pollution is very likely different for both cameras.	Two cameras both have considerable comparable amount of dirt/pollution but with different distributions
933	Obs. Opto.	Before	Transparency	Shutter opens or closes before this is expected	Photoelectric events are exposed to light out of schedule	Rolling Shutter is causing artifacts which are misinterpreted as object properties	Images contain rolling shutter artifacts (both cameras are triggered at the same time but moving objects get distorted due to the rolling shutter)
955	Obs. Opto.	Where else	Spectrum	Different parts of spectrum are transmitted to different locations (chromatic aberration)	Washed-out/Defocused edges	Stereo imaging: matching preciseness decreased	Scene with considerable chromatic aberration and many visible edges
961	Obs. Opto.	More	Lenses number	More lenses are in lens assembly than expected	More reflections between lenses	Lens reflections are misinterpreted as textures or objects	Lens creates double images of parts from the scenery
982	Obs. Opto.	More	Lenses geometry	More optical effects due to strongly curved lens surfaces	Distortion: barrel or pincushion	Distortion: scene geometry misinterpreted	Image with radial distortion not perfectly removed (e.g. somewhat bad intrinsic calibration)
983	Obs. Opto.	More	Lenses geometry	More optical effects due to strongly curved lens surfaces	Vignetting: image darkening toward the edges	Vignetting: increased	Images have considerably amounts of vignetting and scene contains many objects close to the observer
989	Obs. Opto.	Less	Lenses geometry	Less optical corrections due to weakly curved lens surfaces	Focus range is limited in distance (long-sighted)	Close objects defocused - poorly recognized	Scene contains some sharp parts in the background and increasingly out-of-focus parts in the foreground

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
998	Obs. Opto.	Spatial aperiodic	Lenses geom.	Spatial aper. disturbance or imperfections of lens geometry	Bright rays virtually emanate from bright objects within scene	Objects in defect zones of image are not detected correctly	Scratches or rain drops in front of the lens create long bright streaks emanating from all l. s.s in the scene (lens flare?)
1016	Obs. Opto.	Less	Focusing	DoF is smaller than expected	Essential scene parts are out of focus	Blurred image areas misinterpreted as being empty or "medium only"	Images background and main objects in the scene are out of focus
1059	Obs. Opto.	Where else	Aperture	Aperture form is projected into different places within the image	Chromatic aberration in shape of aperture (See More Colour)	Aperture projection is mistaken for an object	Bokeh is visible on the image and has a shape and position to make it prone to confusions with other parts of the image. Critical case for stereo vision: Bokeh and confusion object lie on the same epipolar line
1090	Obs. Opto.	No	Optical Point Spread Function	No optical blurring before discretisation	Staircasing of edges and lines, Aliasing artifacts	Apparent texture differs from true texture	Image contains strong aliasing artifacts
1091	Obs. Opto.	No	Opt. PSF	No optical blurring before discretisation	Moire patterns in intensity and colour of repetitive textures	Unpredictable differences between appearance of corresponding points/regions	Very different textures in left and right image due to large scale Moire effects
1094	Obs. Opto.	More	Opt. PSF	PSF's extent is larger than expected	Loss of contrast, PSF effects a bigger neighbourhood of pixels	Loss of small objects	One of the two sensors is somewhat out of focus
1105	Obs. Opto.	Spatial periodic	Opt. PSF	Periodic pattern visible in the PSF, *i.e.* The PSF is spatially periodic	Additional small scale blurriness creating a spatial pattern	Contours of objects are duplicated and create possibility for confusions	Inter-lens reflections create visible copy of objects in the image

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
1120	Obs. Electr.	More	Exp./Shutter	Longer exposure time than expected	More light captured per image than expected	Overexposure	One of the two images is largely overexposed while the other still shows a lot of detail
1123	Obs. Electr.	Less	Exp./Shutter	Shorter exposure time than expected	Less light captured per image than expected	Underexposure	Large image area is underexposed and shows a lot of blacklevel noise there
1126	Obs. Electr.	As well as	Exposure and shutter	Multiple exposures	Multiple frames superimposed into one image	Movement is miscalculated	Two previous frames are blended/combined into one image
1162	Obs. Electr.	Part of	Resolution (spatial)	Only along one dimension Resol. is different from expected	Only along one dimension Resol. is different from expected	Image/pixel ratio other than expected leading to image distortions	Image before rectification originates from considerably rectangular pixels (instead of square, near to e.g. 2:1 ratio)
1166	Obs. Electr.	Part of	Resolution (spatial)	Part of pixel area is insensitive	Part of pixel area is insensitive	Noise increased	Images contain strong static image noise for well-lit scenes
1168	Obs. Electr.	Reverse	Resolution (spatial)	Resolution is n*m instead of m*n	Resolution is n*m instead of m*n	Size of pixel lines and columns reversed, but number of pixels per image as expected	Image has a considerably larger height than width (untypical image dimensions)
1222	Obs. Electr.	Less	Quality	More overflow effects than expected	E.g. blooming	Blooming effects misinterpreted as objects or object parts	Large difference in light intensity between indoor and outdoor creates large blooming effects around the edges of a window
1261	Obs. Electr.	Reverse	Quantization/Sampling	Recorded intensity is encoded inverse to expected	Image is encoded as its "negative"	Scene recognition breaks down	One camera delivers image negative instead

Continued on next page

Table 1: Full table of specialized hazard list for the stereo vision task

hid	Loc	GW	Par	meaning	consequence	hazard	entry
1265	Obs. Electr.	Other than	Quantization/Sampling	Value Quantisation is other than expected	Intensity output is other than expected	Colours and shadows misinterpreted, derived scene geometry has systematic deviations	Images use logarithmic quantization instead of linear (wrong gamma mapping; mid-tones are washed out)

4. Found Hazard Frames (Accepted as well as Disputed)

Table 2. Accepted Hazard Frames from the Freiburg Dataset

HID	Frame
22	driving/15mm_focallength_scene_forwards_slow/_frame_0298
45	driving/15mm_focallength_scene_forwards_slow/_frame_0200
125	driving/15mm_focallength_scene_forwards_slow/_frame_0064
142	driving/15mm_focallength_scene_forwards_slow/_frame_0652
271	monkaa/a_rain_of_stones_x2/_frame_0051
305	monkaa/family_x2/_frame_0101
326	driving/15mm_focallength_scene_forwards_slow/_frame_0165
476	driving/15mm_focallength_scene_backwards_slow/_frame_0475
481	driving/15mm_focallength_scene_forwards_slow/_frame_0697
561	monkaa/treeflight_x2/_frame_0187
651	driving/35mm_focallength_scene_backwards_slow/_frame_0751
729	driving/35mm_focallength_scene_forwards_slow/_frame_0073
899	monkaa/top_view_x2/_frame_0071
904	driving/15mm_focallength_scene_backwards_slow/_frame_0487
1016	monkaa/funnyworld_camera2_augmented0_x2/_frame_0045
1090	monkaa/treeflight_augmented0_x2/_frame_0100

Table 3. Accepted Hazard Frames from the HCI Dataset

HID	Frame
7	0_0065_frame_02584
22	0_0038_frame_03216
45	0_0065_frame_02480
46	0_0068_frame_02652
47	0_0068_frame_02636
52	0_0026_frame_02296
125	0_0059_frame_04656
141	0_0038_frame_03784
142	0_0038_frame_03488
244	0_0013_frame_08000
444	0_0026_frame_02312
539	1_0014_frame_03408
555	1_0075_frame_05436
666	0_0026_frame_01992
701	0_0068_frame_02492
922	1_0075_frame_05044

Table 4. Accepted Hazard Frames from the KITTI Datasets

HID	Frame
0	kitti2015_000104_frame_10
26	kitti2012_000071_frame_10
50	kitti2015_000144_frame_10
125	kitti2012_000116_frame_10
141	kitti2015_000104_frame_10
142	kitti2012_000120_frame_10
459	kitti2012_000026_frame_10
482	kitti2012_000051_frame_10
651	kitti2012_000010_frame_10
655	kitti2012_000191_frame_10
666	kitti2012_000136_frame_10
701	kitti2012_000193_frame_10
904	kitti2012_000097_frame_10
922	kitti2012_000074_frame_10

Table 5. Accepted Hazard Frames from the Middlebury Datasets

HID	Frame
22	middl_2014_add_Classroom1_perfect_frame_L3_E6
50	middl_2006_orig_Midd1_frame_illum_2_expo_2
52	middl_2006_orig_Monopoly_frame_illum_3_expo_2
125	middl_2006_orig_Midd2_frame_illum_3_expo_1
376	middl_2014_train_orig_Recycle_frame_0
444	middl_2006_orig_Plastic_frame_illum_2_expo_1
449	middl_2005_orig_Laundry_frame_illum_2_expo_2
451	middl_2014_train_orig_Pipes_frame_0
476	middl_2014_train_orig_Jadeplant_frame_0
482	middl_2014_train_orig_Vintage_frame_0
608	middl_2014_train_orig_Jadeplant_frame_0
626	middl_2014_add_Sword2_perfect_frame_L0_E3
735	middl_2005_orig_Laundry_frame_illum_3_expo_1
892	middl_2014_train_orig_PlaytableP_frame_0

Table 6. Accepted Hazard Frames from the Sintel Dataset

HID	Frame
52	sleeping_1_frame_0050
141	shaman_3_frame_0001
142	market_5_frame_0002
183	mountain_1_frame_0031
305	ambush_2_frame_0004
321	ambush_2_frame_0014
326	ambush_2_frame_0012
365	ambush_4_frame_0011
449	shaman_3_frame_0032
459	market_6_frame_0004
539	bamboo_2_frame_0011
883	bandage_2_frame_0011
898	bandage_2_frame_0011
899	mountain_1_frame_0050
904	ambush_2_frame_0004
989	ambush_2_frame_0014

Table 7. Disputed Hazard Frames from the Freiburg Dataset

HID	Frame
52	driving/15mm_focallength_scene_backwards_slow/_frame_0133
141	driving/15mm_focallength_scene_backwards_slow/_frame_0796
259	monkaa/lonetree_augmented1_x2/_frame_0446
321	driving/35mm_focallength_scene_forwards_slow/_frame_0081
383	driving/15mm_focallength_scene_backwards_slow/_frame_0140
555	monkaa/treeflight_x2/_frame_0017
671	driving/15mm_focallength_scene_backwards_slow/_frame_0615

Table 8. Disputed Hazard Frames from the HCI Dataset

HID	Frame
6	0_0038_frame_03216
21	0_0038_frame_03720
47	0_0068_frame_02796
321	1_0026_frame_02760
326	0_0026_frame_02328
376	0_0013_frame_08392
383	0_0050_frame_08696
449	0_00_23_frame_02500
451	0_0000_frame_04816
457	0_0013_frame_08392
459	0_0050_frame_08600
478	0_0026_frame_02272
481	0_0000_frame_04904
482	0_0038_frame_03224
502	0_0013_frame_07968
542	1_0026_frame_02616
561	1_0014_frame_03352
608	0_0000_frame_04400
626	0_0059_frame_09972
655	0_0068_frame_02476
688	0_0038_frame_03728
693	0_0038_frame_03424
698	0_0000_frame_04936
701	0_0026_frame_02304
707	0_0068_frame_02580
729	0_0026_frame_02152
748	1_0032_frame_02364
899	0_0067_frame_03900
998	0_0038_frame_03536
1059	0_0065_frame_02504

Table 9. Disputed Hazard Frames from the KITTI Datasets

HID	Frame
21	kitti2012_000071_frame_10
26	kitti2015_000061_frame_10
46	kitti2012_000074_frame_10
50	kitti2012_000020_frame_10
459	kitti2015_000088_frame_10
509	kitti2015_000058_frame_10
539	kitti2012_000116_frame_10
586	kitti2012_000193_frame_10
655	kitti2015_000169_frame_10
671	kitti2012_000143_frame_10
707	kitti2012_000193_frame_10
922	kitti2015_000062_frame_10

Table 10. Disputed Hazard Frames from the Middlebury Datasets

HID	Frame
0	middl_2014_add_Cable_perfect_frame_L1_E7
275	middl_2006_orig_Bowling1_frame_illum_3_expo_1
383	middl_2001_orig_tsukuba_frame_3
449	middl_2014_train_orig_Playtable_frame_0
457	middl_2014_train_orig_Playtable_frame_0
458	middl_2006_orig_Wood2_frame_illum_2_expo_1
481	middl_2006_orig_Bowling1_frame_illum_2_expo_0
482	middl_2005_orig_Laundry_frame_illum_2_expo_1
539	middl_2014_train_orig_Playtable_frame_0
883	middl_2001_orig_map_frame_0
1123	middl_2006_orig_Baby3_frame_illum_1_expo_0

Table 11. Disputed Hazard Frames from the Sintel Dataset

HID	Frame
244	bamboo_2_frame_0007
275	bandage_1_frame_0028
451	market_6_frame_0005
481	ambush_6_0003
1123	shaman_3_frame_0050

Table 12. Summary of datasets and the associated url.

NAME	YEAR	URL
Middlebury	2002	http://vision.middlebury.edu/stereo/data/scenes2001/
Middlebury	2003	http://vision.middlebury.edu/stereo/data/scenes2003/
Middlebury	2007	http://vision.middlebury.edu/stereo/data/scenes2006/
EISATS S1	2008	http://ccv.wordpress.fos.auckland.ac.nz/eisats/set-1/
EISATS S2	2008	http://ccv.wordpress.fos.auckland.ac.nz/eisats/set-2/
EISATS S6	2009	http://ccv.wordpress.fos.auckland.ac.nz/eisats/set-6/
New College	2009	http://www.robots.ox.ac.uk/NewCollegeData/
Pittsburgh	2009	http://pfid.rit.albany.edu/
EVD	2011	http://cmp.felk.cvut.cz/wbs/#datasets
Ford Campus	2011	http://robots.engin.umich.edu/SoftwareData/Ford
HCI-Robust	2012	https://hci.iwr.uni-heidelberg.de/Robust_Vision_Challenge_2012
KITTI 2012	2012	http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo
Leuven	2012	https://www.inf.ethz.ch/personal/ladickyl/Leuven.zip
Tsukuba	2012	http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php
HCI-Synth	2013	http://heidata.customers.aldago.com/dataset
Stixel	2013	http://www.6d-vision.com/ground-truth-stixel-dataset
Daimler Urban	2014	http://www.6d-vision.com/scene-labeling
Malaga Urban	2014	http://www.mrpt.org/MalagaUrbanDataset
Middlebury	2014	http://vision.middlebury.edu/stereo/data/scenes2014/
Cityscapes	2015	https://www.cityscapes-dataset.com/
KITTI 2015	2015	http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo
MPI Sintel	2015	http://sintel.is.tue.mpg.de/stereo
Freiburg CNN	2016	http://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html
HCI Training	2016	http://www.hci-benchmark.org/dataset
SYNTHIA	2016	http://synthia-dataset.net/
Virtual KITTI	2016	http://www.xrce.xerox.com/Research-Development/Computer-Vision/Proxy-Virtual-Worlds
Oxford RobotCar	To appear	http://robotcar-dataset.robots.ox.ac.uk/

4.2 Unifying Panoptic Segmentation for Autonomous Driving

Supplemental Material CVPR2022 Paper
Unifying Panoptic Segmentation for Autonomous Driving

1. Unified Label Policy

Table 1 lists all WD2 labels including mappings to MVD, IDD, Cityscapes, and $WD2_{eval}$ as well as each label's supercategory and visualization color.

	WD2	color	MVD	IDD	Cityscapes	$WD2_{eval}$
human	**person**		**person**	**person**	**person**	**person**
human	**motorcyclist**		**motorcyclist**	**rider**	**rider**	**rider**
human	**bicyclist**		**bicyclist**	**rider**	**rider**	**rider**
human	**otherrider**		**otherrider**	**rider**	**rider**	**rider**
vehicle	**egovehicle**		egovehicle	**egovehicle**	*egovehicle*	**egovehicle**
vehicle	dashcammount		carmount	**egovehicle**	*egovehicle*	**egovehicle**
vehicle	**car**		**car**	**car**	**car**	**car**
vehicle	**truck**		**truck**	**truck**	**truck**	**truck**
vehicle	**bus**		**bus**	**bus**	**bus**	**bus**
vehicle	**motorcycle**		**motorcycle**	**motorcycle**	**motorcycle**	**motorcycle**
vehicle	**bicycle**		**bicycle**	**bicycle**	**bicycle**	**bicycle**
vehicle	**pickup**		**truck**	**truck**	**truck**	**pickup**
vehicle	**van**		**car**	**car**	**car**	**van**
vehicle	**autorickshaw**		**othervehicle**	**autorickshaw**	**motorcycle**	**motorcycle**
vehicle	**caravan**		**caravan**	**caravan**	***caravan***	*unlabeled*
vehicle	**trailer**		**trailer**	**trailer**	***trailer***	*unlabeled*
vehicle	**onrails**		onrails	**train**	**onrails**	*unlabeled*
vehicle	**othervehicle**		**othervehicle**	**vehiclef.**	*dynamic*	*unlabeled*
vehicle	**wheeledslow**		**wheeledslow**	**vehiclef.**	*dynamic*	*unlabeled*
vehicle	**boat**		**boat**	*unlabeled*	*dynamic*	*unlabeled*
flat	road		road	road	road	road
flat	sidewalk		sidewalk	sidewalk	sidewalk	sidewalk
flat	roadmarking		markinggeneral	road	road	roadmarking
flat	curb		curb	curb	sidewalk	sidewalk
flat	tramtrack		railtrack	road	road	road
flat	bikelane		bikelane	road	road	road
flat	bikelanesidewalk		bikelane	sidewalk	sidewalk	sidewalk
flat	pedestrianarea		pedestrianarea	road	road	road
flat	**crosswalkplain**		**crosswalkplain**	road	road	road
flat	**crosswalkzebra**		**crosswalkzebra**	road	road	road
flat	curbterrain		curb	curb	terrain	terrain
flat	servicelane		servicelane	road	road	road
flat	curbcut		curbcut	curb	sidewalk	sidewalk
flat	ground		*unlabeled*	*unlabeled*	*ground*	*unlabeled*
flat	parking		parking	parking	*parking*	*unlabeled*
flat	railtrack		railtrack	railtrack	*railtrack*	*unlabeled*
flat	water		water	nondrivablef.	*ground*	*unlabeled*
flat	sand		sand	drivablef.	*ground*	*unlabeled*
flat	snow		snow	*unlabeled*	*ground*	*unlabeled*
object	polegroup		**pole**	*polegroup*	*polegroup*	pole
object	**pole**		**pole**	pole	*pole*	pole
object	**utilitypole**		**utilitypole**	pole	*pole*	pole
object	**trafficsignframe**		**trafficsignframe**	pole	*pole*	pole
object	**trafficlight**		**trafficlight**	trafficlight	trafficlight	trafficlight
object	**billboard**		**billboard**	billboard	*billboard*	billboard
object	**streetlight**		**streetlight**	obsstrbarf.	*streetlight*	streetlight
object	**manhole**		**manhole**	road	road	road
object	**trafficsign**		**trafficsignfront**	trafficsign	trafficsign	trafficsign
object	**trafficsignback**		**trafficsignback**	obsstrbarf.	*static*	*unlabeled*
object	**trafficsignany**		*unlabeld*	obsstrbarf.	*static*	*unlabeled*
object	otherbarrier		otherbarrier	wall	wall	wall
object	**catchbasin**		**catchbasin**	road	road	road
object	**manholesidewalk**		**manhole**	sidewalk	sidewalk	sidewalk
object	**junctionbox**		**junctionbox**	obsstrbarf.	*static*	*unlabeled*
object	**mailbox**		**mailbox**	obsstrbarf.	*static*	*unlabeled*
object	**phonebooth**		**phonebooth**	obsstrbarf.	*static*	*unlabeled*
object	**bikerack**		**bikerack**	obsstrbarf.	*static*	*unlabeled*
object	pothole		pothole	road	road	road
object	**trashcan**		**trashcan**	obsstrbarf.	*static*	*unlabeled*
object	**bench**		**bench**	obsstrbarf.	*static*	*unlabeled*
object	**banner**		**banner**	obsstrbarf.	*dynamic*	*unlabeled*
object	**firehydrant**		**firehydrant**	obsstrbarf.	*static*	*unlabeled*
object	**cctvcamera**		**cctvcamera**	obsstrbarf.	*static*	*unlabeled*
construction	building		building	building	building	building
construction	wall		wall	wall	wall	wall
construction	fence		fence	fence	fence	fence
construction	guardrail		guardrail	guardrail	*guardrail*	guardrail
construction	bridge		bridge	bridge	*bridge*	*unlabeled*
construction	tunnel		tunnel	tunnel	*tunnel*	*unlabeled*
nature	vegetation		vegetation	vegetation	vegetation	vegetation
nature	terrain		terrain	nondrivablef.	terrain	terrain
nature	**groundanimal**		**groundanimal**	**animal**	*dynamic*	*unlabeled*
nature	**bird**		**bird**	**animal**	*dynamic*	*unlabeled*
nature	mountain		mountain	f.background	*static*	*unlabeled*
sky	sky		sky	sky	sky	sky
void	dynamic		*unlabeled*	*unlabeled*	*dynamic*	*unlabeled*
void	overlay		*unlabeled*	*rect.border*	*rect.border*	*unlabeled*
void	outofroi		*unlabeled*	*outofroi*	*outofroi*	*unlabeled*
void	static		*unlabeled*	*unlabeled*	*static*	*unlabeled*
void	unlabeled		*unlabeled*	*unlabeled*	*unlabeled*	*unlabeled*

Table 1. Wilddash2 label policy and mapping to MVD, IDD, CS, and $WD2_{eval}$. Bold labels have instance annotations, italic labels are not evaluated at their respective benchmark. Negative test cases do evaluate areas labeled as *unlabeled* in $WD2_{eval}$ (see paper's Section 4.2 on Negative Testing); Supercategories: human; vehicle; flat; object; construction; nature; sky; void

2. Category definitions

The WD2 label policy unifies MVD, IDD, and Cityscapes category labels (in addition to the new vehicle labels *pickup* and *van*. The definition for most labels can be found in existing label definitions. Others need clarification or clear rules for differentiation in borderline cases. This leads to the following category definitions:

- The categories *person*, *egovehicle*, *car*, *truck*, *bus*, *motorcycle*, *bicycle*, *caravan*, *trailer*, *onrails*, *road*, *sidewalk*, *ground*, *parking*, *railtrack*, *polegroup*, *billboard*, *streetlight*, *building*, *wall*, *fence*, *guardrail*, *bridge*, *tunnel*, *vegetation*, *terrain*, *sky*, *unlabeled*, *outofroi*, *static*, and *dynamic* are described in the supplemental material to the Cityscapes [1] paper.
- The categories *motorcyclist*, *bicyclist*, *otherrider*, *othervehicle*, *wheeledslow*, *boat*, *roadmarking* (==*marking general*) *curb*, *bikelane*, *pedestrianarea*, *crosswalkplain*, *crosswalkzebra*, *servicelane*, *curbcut*, *water*, *sand*, *snow*, *pole*, *utilitypole*, *trafficlight*, *trafficsign* (== *traffic sign front*), *manhole*, *pothole*, *trafficsignback*, *trafficsignframe*, *otherbarrier*, *catchbasin*, *junctionbox*, *mailbox*, *phonebooth*, *bikerack*, *trashcan*, *bench*, *banner*, *firehydrant*, *cctvcamera*, *groundanimal*, *bird*, *mountain*, *dashcammount* (== *car mount*) are described in the supplemental material to the MVD [2] paper.

	human	vehicle	flat	object	construction	nature	sky	average
mvd100 PQ_{Cat}	46.0%	55.3%	71.6%	32.3%	52.8%	66.5%	79.5%	57.7%
mix150 PQ_{Cat}	49.7%	61.4%	86.2%	34.1%	62.5%	71.9%	87.3%	64.7%
mvd100 RQ_{Cat}	60.2%	67.0%	84.1%	48.2%	70.0%	82.1%	86.3%	71.1%
mix150 RQ_{Cat}	65.2%	73.6%	97.5%	50.9%	79.8%	87.1%	93.6%	78.2%
mvd100 SQ_{Cat}	76.4%	82.6%	85.1%	67.1%	75.4%	80.9%	92.2%	80.0%
mix150 SQ_{Cat}	76.2%	83.4%	88.4%	66.9%	78.2%	82.6%	93.3%	81.3%

Table 2. Per-supercategory PQ, RQ and SQ metrics evaluated on the hidden WD2 benchmark set for both models *mvd100* and *mix150* presented in the main paper.

- The labels *curb* and *curbterrain* both are described by the MVD *curb* label (i.e. curb stones; including all visible faces of a curb). If the curb encases an area labeled with terrain (or other vegetation), then the curb receives the *curbterrain* label. Otherwise use *curb*.
- The labels *bikelane* and *bikelanesidewalk* are both described by the MVD *curb* label. Use *bikelanesidewalk* if the bikelane is on a sidewalk. Otherwise use *bikelane*.
- The labels *manhole* and *manholesidewalk* are both described by the MVD *manhole* label. Use *manholesidewalk* if the manhole is on a sidewalk. Otherwise use *manhole*.
- The *traintrack* is described by the Cityscapes *traintrack* label (i.e. track of raised rails, not drivable by cars). The *tramtrack* label is used for the track area between embedded rails (drivable by cars) including the rails themselves.
- The *autorickshaw* category is described in the IDD [3] paper.
- The *trafficsignany* category is a fallback category used for cases where either *trafficsign* (=front) or *trafficsignback* could be correct.
- Vehicle class *pickup*: This label is used for light commercial vehicles (LCV) with an open cargo area. It only applies to motorized, car-sized vehicles with a visible un-roofed cargo area (also for open cages). Pickup trucks have regular car front wheels and a regular car wheelbase (distance). Cargo vehicles with larger tires or a truck motor housing (driver sitting above motor with a vertical windscreen and bonnet) retain the *truck* label.
- Vehicle class *van*: This label applies to motorized LCV without an open cargo area. Vans have a boxy shape with regular car tires and their wheel-base is typically larger than those of regular cars. The front of vans is often inclined but straight and they are distinctively higher (i.e. the van's ceiling) than regular cars. Vehicles sold under the term "mini-van" with a regular car height remain at the *car* label. Hybrid vehicles with a fully separated, non-continuous, driver cabin are still labeled as *truck* (e.g. many ambulance vehicles, police, some delivery trucks).

3. Supercategory Scores

Table 2 shows individual per-supercategory scores of mvd100 and mix150 on the $WD2_bench$ set. Table 3 in the main paper contains the right-most column (arithmetic mean over all supercategories) denoted as PQ_{cat}.

4. Further Experiments

The following Table 3 mirrors the layout of Table 3 in the main paper. Experiments were conducted for WD2, Cityscapes, and IDD in the same way as for MVD: first 100 epochs on only the original dataset using the standard training label policy (plus *van* and *pickup*. This results in 66 labels for MVD, 29 for IDD, and 22 for Cityscapes. The validation results are calculated on the original validation dataset using the original training label policy while the WD2 benchmark evaluation remaps the algorithm outputs to the $WD2_{eval}$ label policy (26 labels) and reports the results for the hidden WD2 benchmark dataset. The models after 100 epochs are fine-tuned for 50 epochs using WD2 individually remapped to each of the dataset's training label policies mixed with the same amount of frames from the original dataset (i.e. 50%/50% split). The model for the first column *wd2_100* uses only WD2 frames with the $WD2_{eval}$ label policy during training. The same train/val split is used both for *wd2_100* as well as in all fine-tunings.

	Original Validation					*WD2* Benchmark						
	PQ	SQ	RQ	PQ_{van}	PQ_{pickup}	PQ	SQ	RQ	PQ_{van}	PQ_{pickup}	PQ_{neg}	PQ_{cat}
wd2_100	38.0%	75.6%	48.2%	36.9%	33.4%	37.0%	75.5%	47.7%	35.1%	37.3%	16.9%	61.1%
cs100	55.7%	76.4%	68.2%	53.7%	0.0%	10.7%	69.5%	15.0%	10.6%	0.0%	5.4%	22.8%
cs150	56.1%	77.4%	68.2%	55.2%	0.0%	32.2%	76.9%	41.1%	34.4%	41.1%	15.5%	58.4%
idd100	47.7%	75.6%	59.5%	48.8%	0.0%	15.3%	72.8%	20.1%	14.1%	0.0%	7.2%	35.7%
idd150	46.4%	75.7%	57.9%	40.4%	0.0%	29.4%	76.1%	37.7%	33.7%	33.7%	14.2%	54.8%
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. Comparison of performances for multiple models first trained on the respective original datasets (WD2, Cityscapes, IDD, MVD) and later fine-tuned for 50 additional epochs on a 50%/50% mixture of the original dataset and WD2. The left side shows results when evaluated on the original validation sets and the right side shows scores on the hidden WD2 benchmark set. The label policy on the left is not constant across the rows while the right side all evaluate using the same labels: $WD2_{eval}$. Results for MVD are duplicated from the main paper to improve comparability.

References

[1] 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. 1

[2] 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. 1

[3] 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. 2

Bibliography

[BSL+11] Simon Baker, Daniel Scharstein, J. Lewis, Stefan Roth, Michael Black, and Richard Szeliski. A database and evaluation methodology for optical flow. 92:1–31, 2011.

[BWSB12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. 2012.

[Cem03] Kaner Cem. What is a good test case. In *Software Testing Analysis & Review Conference (STAR) East*, 2003.

[COR+16a] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[COR+16b] 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. 2016.

[CVH22] Cv-hazop - vitro. `https://vitro-testing.com/cv-hazop/`, 2022. Accessed: 2022-11-22.

[DCS+17] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017.

[GLGL18] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012.

[Ike21] Katsushi Ikeuchi. *Computer vision: A reference guide*. Springer, 2021.

[imp23a] Best computer science conferences - ranking 2022. https://research.com/conference-rankings/computer-science, 2023. Accessed: 2023-04-20.

[imp23b] Journal rankings on computer vision and pattern recognition. https://www.scimagojr.com/journalrank.php?category=1707, 2023. Accessed: 2023-04-20.

[JB19] Richard A Johnson and Gouri K Bhattacharyya. *Statistics: principles and methods*. John Wiley & Sons, 2019.

[JGB+20] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020.

[KHG+19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[Kle83] Trevor A Kletz. *Hazop & Hazan: Hazard Workshop Modules: Notes on the Identification and Assessment of Hazard*. Institution of Chemical Engineers, 1983.

[KNH+16] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, and Bernd Jahne. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. 2016.

[KRA+20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Mat11] Stefano Mattoccia. Stereo vision: Algorithms and applications. *University of Bologna*, 22, 2011.

[MG15] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. 2015.

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[NORBK17] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[RVC22] Robust vision challenge 2022. http://www.robustvision.net/, 2022. Accessed: 2022-11-22.

[RVRK16] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. 2016.

[SCD+06] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. 2006.

[Sch11] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social choice and Welfare*, 36(2):267–303, 2011.

[sch23] Oliver zendel - google scholar. `https://scholar.google.com/citations?user=7mztkLgAAAAJ`, 2023. Accessed: 2023-04-20.

[Sem12] Godwin Sebabi Semwezi. Automation of negative testing. 2012.

[SGB+20] Hendrik Schilling, Marcel Gutsche, Alexander Brock, Dane Spath, Carsten Rother, and Karsten Krispin. Mind the gap-a benchmark for dense depth prediction beyond lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 338–339, 2020.

[SHJ11] Rupert Schlick, Wolfgang Herzner, and Elisabeth Jöbstl. Fault-based generation of test cases from uml-models–approach and some experiences. In *Computer Safety, Reliability, and Security: 30th International Conference, SAFECOMP 2011, Naples, Italy, September 19-22, 2011. Proceedings 30*, pages 270–283. Springer, 2011.

[SHK+14] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. 2014.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. 47:7–42, 2002.

[SSG+17] Thomas Schöps, Johannes Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. 2017.

[ZHM13] Oliver Zendel, Wolfgang Herzner, and Markus Murschitz. Vitro-model based vision testing for robustness. In *Proceedings of the 44th Internationel Symposium on Robotics, ISR 2013, Seoul, Korea (South), October 24-26, 2013*, pages 1–6. IEEE, 2013.

[ZHM+17] Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernández Domínguez. Analyzing computer vision data - the good, the bad and the ugly. In *Proceedings of the 30th Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, United States of America, July 21-26, 2017*, pages 6670–6680. IEEE, 2017.

[ZHM+18] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *Proceedings of the 15th European Conference on Computer Vision, ECCV 2018, Munich, Germany, September 8-14, 2018*, volume 11210 of *Lecture Notes in Computer Science*, pages 407–421. Springer, 2018.

[ZMHH15] Oliver Zendel, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. CV-HAZOP: introducing test data validation for computer vision. In *Proceedings of the 15th International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2066–2074. IEEE, 2015.

[ZMHH17] Oliver Zendel, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. How good is my test data? introducing safety analysis for computer vision. *International Journal of Computer Vision*, 125(1-3):95–109, 2017.

[ZMZ+19] Oliver Zendel, Markus Murschitz, Marcel Zeilinger, Daniel Steininger, Sara Abbasi, and Csaba Beleznai. Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the 34th Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, United States of America, June 16-20, 2019*, pages 1221–1229. IEEE, 2019.

[ZSR+22] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *Proceedings of the 35th Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, United States of America, June 18-24, 2022*, pages 21319–21328. IEEE, 2022.

[ZZ21] Oliver Zendel and Christian Zinner. Naphash: Efficient image hash to reduce dataset redundancy. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 2021.

[ZZP+17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.