The HCI Benchmark Suite: Stereo And Flow Ground Truth With Uncertainties for Urban Autonomous Driving

Daniel Kondermann* Rahul Nair* Katrin Honauer* Karsten Krispin* Jonas Andrulis†
Alexander Brock* Burkhard Güssefeld* Mohsen Rahimimoghaddam* Sabine Hofmann‡
Claus Brenner‡ Bernd Jähne*

*Heidelberg Collaboratory for Image Processing,
Ruprecht-Karls Universität Heidelberg, Germany
f.l@iwr.uni-heidelberg.de

†Pallas Ludens GmbH,
Heidelberg, Germany
f.l@pallas-ludens.com

‡Institute of Cartography and Geoinformatics,
Leibniz Universität Hannover, Germany
f.l@ikg.uni-hannover.de

Abstract

Recent advances in autonomous driving require more and more highly realistic reference data, even for difficult situations such as low light and bad weather. We present a new stereo and optical flow dataset to complement existing benchmarks. It was specifically designed to be representative for urban autonomous driving, including realistic, systematically varied radiometric and geometric challenges which were previously unavailable.

The accuracy of the ground truth is evaluated based on Monte Carlo simulations yielding full, per-pixel distributions. Interquartile ranges are used as uncertainty measure to create binary masks for arbitrary accuracy thresholds and show that we achieved uncertainties better than those reported for comparable outdoor benchmarks. Binary masks for all dynamically moving regions are supplied with estimated stereo and flow values.

An initial public benchmark dataset of 55 manually selected sequences between 19 and 100 frames long are made available in a dedicated website featuring interactive tools for database search, visualization, comparison and benchmarking.

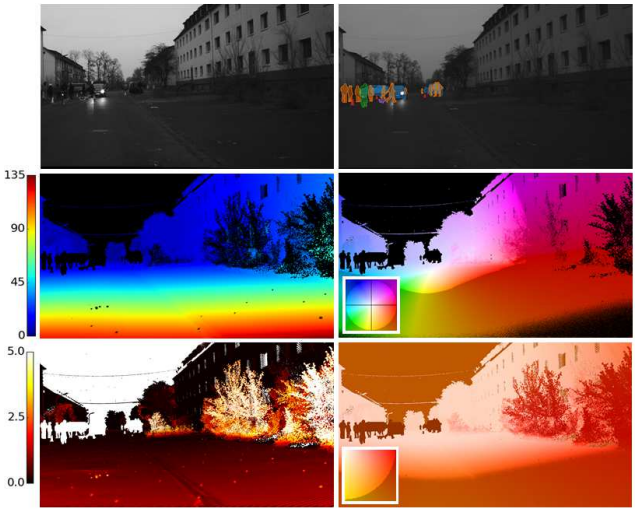

Figure 1. Top: sample image (left) and masks of dynamic regions with labels (right). Center: stereo (left) and flow ground truth (right), masked with certainties better than two pixels (otherwise black). Bottom: uncertainties for stereo (left) and flow (right). The HSV color coding for flow is chosen such that the V component is maximal at a flow magnitude of four pixels; all other regions have a lower V value. Our new dataset covers previously unavailable, difficult light and weather situations and supplies ground truth with uncertainties.

1. Introduction

In computer vision, ground truth generation and performance analysis have received increasing attention in the past years [10, 5, 39, 7]. As a result, recent ground truth databases have successfully pushed the limits of optical flow and stereo estimation. One quickly evolving application with safety relevance is autonomous driving. Here, computer vision results such as optical flow and depth are used to make decisions for steering and velocity control. The generation of reference data for these sensors enables researchers to compare methods with respect to various algorithm properties such as accuracy and robustness. In this paper, we introduce an extensive dataset specifically tailored to complement existing flow and stereo benchmarks for urban autonomous driving. It covers previously unavailable, challenging situations such as low light or rain and comes with pixel-wise uncertainties. The main components of our dataset are visualized in Figure 1.

One of the guiding principles for our database design is a direct consequence from the generalization-specialization

dilemma[1] known from machine learning [21]. It states that an algorithm either works reasonably well for a broad range of applications or it works excellently for a very specific domain [23]. In order to achieve superior performance, stereo and flow algorithms therefore need to be defined such that they perform particularly well in a predefined use-case as for example autonomous driving. To empower algorithm designers to find models generalizing well to the full spectrum of the actual use-case, a dataset tailored to the application needs to be representative. A dataset is *representative* if computed benchmark results allow for a prediction of the performance in the actual application.

Our contribution is threefold: first, we designed a ground truth acquisition method for urban autonomous driving datasets (Section 3), including requirements relevant for this use-case. Second, we present the first radiometrically challenging stereo and flow ground truth dataset with full measurement error distributions containing high resolution (HR), high frame rate (HFR) and high dynamic range (HDR) (Section 4.1). Third, we deploy an interactive web application and SDK with detailed visualization and benchmarking tools including search functionality and state of the art performance metrics (Section 4.3).

2. Related Work

Performance analysis for flow and stereo is addressed from two main directions: First, synthetic images can be created for benchmarks (e.g. with computer graphics). Second, reference results for real images are measured with specialized hardware. A third alternative is to not generate ground truth at all and leave the benchmarking to experts.

Creating synthetic ground truth by simulation is a very flexible approach. It is relatively straightforward to generate flow and depth ground truth while allowing for systematic variations in all scene parameters such as material properties, light sources as well as animations. Important early flow and stereo datasets were [4, 36, 29, 44, 26, 34, 43]. The first widely recognized synthetic dataset with a benchmarking website was MPI-Sintel [5] which used existing assets from a Blender movie to generate a dataset termed *naturalistic*, addressing the fact that the data looks a bit more like a cartoon than the real world, but still resembles real images to some degree [46]. In [35], the focus lies on simulating motion blur for simultaneous localization and mapping, whereas [14] suggests to use computer game engines to generate large amounts of data by simply playing a game. A dataset for learning occlusion boundaries was presented in [18], including published tools to create new datasets.

While some preliminary statistical results indicate that synthetic data can be used as ground truth [46], these methods have not been thoroughly evaluated with respect to their representativeness [31, 12].

Creating ground truth by measurement is based on real-world images. Here the challenge lies in creating reference data which is accurate enough. One option is to record real data and use *manual measurements*. Some success was achieved both with expert [27, 25] as well as laymen annotations [9]. Although the accuracy in [27] is good compared to reference data from the Middlebury flow benchmark, annotations are not measurements. Possible biases introduced by humans have yet to be investigated.

(Semi-)automatic measurement setups have a human in the loop to correct algorithm results. They are more reliable, but only work in restricted scenarios. Proposed methods include using more than two cameras [33] or additional modalities such as structured light [30, 40], LIDAR [10] or UV-paint with multiple exposures and light sources [3]. Another approach is to use approximated GT based on domain specific assumptions, e.g. planar "Stixels" [37, 38],

These approaches are not as costly as completely manual processing. Our dataset follows these approaches in that we use experts to correct ground truth generation algorithm results and obtain as much information as possible from measurement devices. Another downside of such approaches is that they still are prone to outliers and biases caused by measurement devices and human corrections. Yet, it is the only currently known method for large-scale, outdoor stereo and flow ground truth generation.

To deal with such uncertainties during benchmarking, several approaches have been developed. The first Middlebury datasets contain general discussions on accuracy [3]. The latest Middlebury stereo set [39] contains estimated per-pixel standard deviations based on multiple measurements. For LIDAR-based datasets, accuracy based on error propagation was discussed in detail in [42]. A faster method for accuracy estimates based on sampling was presented in [24].

Datasets coming with a benchmark website have become a popular approach to performance analysis. They offer a very diverse in choice of hardware, settings and content. With a benchmark, algorithms can be compared more easily and the research community can focus on relevant challenges in the field. We focus on the four datasets coming with a benchmark most relevant for autonomous driving: KITTI [10], MPI-Sintel [5], Middlebury [39] and Cityscapes [7].

We compared these datasets with respect to the requirements to be discussed in Section 3.1. An overview is given in Figure 5, while details are discussed in Section 4.2. A more comprehensive summary of all dataset properties such as images numbers, resolution and camera settings are given in the supplemental material.

With respect to benchmarking metrics, all current websites are largely based on average, standard deviation, quan-

[1] also referred to as *bias-variance trade-off*

tiles and pixel counts with error thresholds of performance metrics such as disparity and endpoint error. They further come with binary masks containing undefined regions due to occlusions or motion outside the frame. Recent research showed that additional, geometrically meaningful metrics can be used to further describe algorithm properties in the stereo domain [16]. In our benchmarking website, we build on these existing metrics.

3. Methods

Our dataset generation approach comprises three major steps: first, we derive requirements for a representative dataset (Section 3.1). Second, we built a ground truth acquisition system (Section 3.2). Third, we devised a recording strategy to meet the requirements (Section 3.3).

3.1. Dataset Requirements

The goal in autonomous driving is to at least achieve the average reliability of humans [45]. This requires extensive context knowledge such as speed limits and three-dimensional trajectories of traffic participants. More generally, context can be established by exploring all properties of the scene sensed by the car. These can be described by distance, motion, relation and type of all objects. Scene understanding and object recognition describe relation and type of objects. To complement these relatively orthogonal fields of research coming with their own benchmarking methods (e.g. [7, 8]), we focus on distance and motion measurement.

Many autonomous driving systems are based on visual data acquired by cameras. They need to remain reliable whenever images deteriorate due to environmental situations. Typically, stereo and optical flow methods struggle with a number of effects caused by geometric and radiometric challenges such as complex occlusions, fast motion, brightness changes, lens flares, etc. [41].

Based on these observations, we define the first group of content requirements as follows. Note that we do not formulate explicit requirements on traffic rules implying scene understanding based on e.g. road markings and signs. Since the present version of our dataset focuses on dense correspondences, we are less interested in these semantics of a scene and focus more on geometric and radiometric challenges occurring in our scenario. A good overview on scene understanding content requirements can be found in [7].

We require a dataset for urban autonomous driving system to challenge: (R1) robustness against radiometric challenges such as direct sunlight, strong specularities, lens-flares, low light occurring at night, in tunnels or parking lots; (R2) robustness to imaging distractions such as raindrops on the windshield, reflecting puddles on the road, snow and fog; (R3) robustness to changes induced by the time of year (vegetation, pedestrian clothes and sun position) and (R4) robustness with respect to location (longitude/latitude on earth, country, culture and traffic laws). Finally, the behavior of traffic participants should be representative. For low-level vision this amounts to geometric complexity (R5).

To create a representative dataset we need to make sure that (1) the important real-world observations for our use-case are covered by the dataset and that (2) the dataset provides a high enough number of observations so that noise may not lead to overfitting. While (1) calls for a carefully selected bias-free scene content (2) calls for a high number of sequences. We require quality and quantity of the dataset sequences to at least match current best practices in performance analysis (R6).

In addition to the content of the scenes, we need to define a ground truth acquisition system. Our goal is to benchmark the state of the art in stereo and flow for urban autonomous driving for the next years. We assume that currently too expensive systems will soon be built-in parts of next-generation vehicles. Since camera systems are quickly evolving, we require our camera system to be at least as capable as the best commercially available system in terms of high resolution (R7), dynamic range (R8) and frame rate (R9). R7 will further push algorithm development as current methods often cannot handle large images sizes. On the other hand, future algorithms should be able to assume relatively small motions due to R9 and good image quality due to R8. Finally, the ground truth coming with the images should be at least an order of magnitude more accurate than the best available algorithm (R10).

Among the most important results of a recent CVPR workshop on performance analysis[2] was a consensus that the user interface and the way the performance metrics are presented play a crucial role in how researchers use and interpret the data. The attendees further found that simply using a single metric and trying to be at the first rank is not always the best way to design and compare algorithms. Therefore, and to attract many researchers to use the dataset, it should not only be highly accessible (R11) but also allow for state-of-the art comparative performance analysis as well as scientific dissemination (R12).

3.2. Ground Truth With Uncertainties

We use a hardware setup similar to [24][3]. In order to compute depths as well as optical flow based on measurement devices, we used one of the most accurate LIDAR scanners suitable for scanning large-scale outdoor areas. We scanned the empty scene first, excluding all traffic par-

[2] http://hci.iwr.uni-heidelberg.de//Static/cvpr15ws-correspondence/

[3] All details on the LIDAR system are given in the supplemental material.

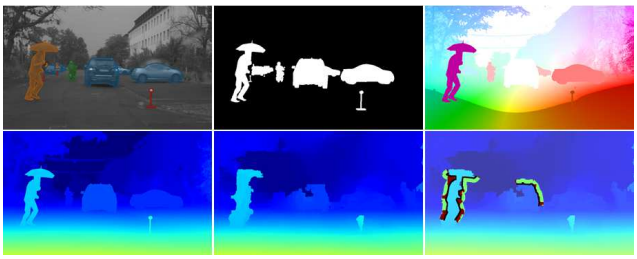

Figure 2. Top: For each frame, we provide individual labels for dynamic objects (left), binary object masks (center) and approximate constant flow vectors per object (right). Bottom: We further provide approximated constant disparity estimates per object (left) which can be used to assess stereo algorithm results (center) for edge fattening (right) and further metrics which do not require pixelwise accurate disparities.

ticipants such as vehicles, pedestrians and movable carry-on items. The average distance between two points in the final point cloud is 1-2 cm on a planar surface, yielding a good density at all relevant locations. This allows for high-accuracy ground truth of all static parts of the scene which are visible to the LIDAR system (cf. R10).

In order to meet imaging quality requirements R6-8, we designed a stereo camera system consisting of two *pco.edge 5.5* cameras with parallel mounted optical axes. The baseline of 30 cm is oriented horizontally, which is commonly used in automotive applications. The system was mounted at rear mirror height behind the windshield of the test vehicle.

The sensor of each camera has the dimensions of $14.04\,\text{mm} \times 16.64\,\text{mm}$ at a resolution of $2160\,\text{px} \times 2560\,\text{px}$ with a pixel pitch of $6.5\mu\text{m}$. We used a configuration with a resolution of $1080\,\text{px} \times 2560\,\text{px}$, so we could achieve a maximal frame rate of almost 200 Hz at horizontal and vertical fields of views of about 70 and 30 degrees respectively. These cameras have a dynamic range of 27000:1 linearly encoded in 16 bit values, which were mapped to 8 bit values through a non-linear noise equilibrating transform without loss of information [19, 20]. Two *Kowa LM12XC* C-Mount lenses with a focal length of 12 mm were used at f-Numbers ranging from 4 to 8, depending on the light conditions. The exposure time was adjusted manually between sequence recordings to avoid saturation in the images and ranged from 0.5 to 4.9 ms.

We determined the internal camera parameters of the stereo camera pair using the method described by Abraham and Hau [1]. The RMS re-projection error reached 0.22 px with a variance of 3.6 px. We measured lens distortions and performed all subsequent calculations on the undistorted and stereo rectified images.

To meet the technical requirements to reduce overfitting due to limited or unknown dataset accuracy (R6) and to further validate the accuracy of the ground truth (R10) we follow the goal of [24] to generate ground truth with uncertainties: we want to know the interquartile range of the disparity and flow at each ground truth pixel. This quantity should be derived from first principles, using all observable error distributions such as those of LIDAR measurements, intrinsic camera calibration estimates and 2D feature correspondence accuracy.

Since in our case the scene is scanned once before the actual recordings with actors, the main task for ground truth generation for stereo and flow lies in registering each stereo image pair with the LIDAR point cloud. This yields a camera pose which can be used to project all 3D points into the 2D image plane. We used the approach described in [24] with few extensions, where features in the LIDAR point cloud where labeled manually and then semi-automatically associated with 2D feature tracks in the stereo images.

Once camera pose and the newly established 2D to 3D feature correspondences in each frame have been found, a final pose estimation is computed using bundle adjustment. The objective function is the same as in [24], yielding optimal camera poses given all known uncertainties in camera calibration, 3D point accuracy and 2D feature accuracy. The final depth and optical flow ground truth as well as the respective pixel-wise error distributions for each frame are computed based on Monte Carlo Sampling as described in [24].

This process yields ground truth with uncertainties for all parts in the sequence which are not individually moving, which this amounts to about 94% of all pixels minus sky regions. In a next step, we manually annotated all dynamically moving regions in all images at 25 Hz temporal resolution with pixel-accurate contours. Each object instance was labeled individually (even if overlapping), allowing for additional labels for future benchmarks e.g. in action recognition. Example annotations can be found in Figure 2. From the polygon contours we created binary masks in which no ground truth is available. In these regions, we provide rough estimates on displacements. To this point, no reliable ground truth exists for these regions. Yet, these regions allow for more qualitative evaluations such as foreground fattening or thinning.

3.3. Recording Strategy

Most current datasets in the automotive domain focus on unconstrained scenes recorded in the public. This comes with the advantage that the real-world distribution can be sampled relatively uniformly, assuming that biases due to recording in e.g. a single country are negligible. A disadvantage is that rarely occurring events such as accidents are hard to acquire, resulting in a reduced representativeness.

In order to model all requirements addressing the representativeness of content (R1-5) we were challenged with the combinatorial explosion of possible environmental and

Figure 3. A Google Maps satellite photo of the location where recording took place. The part of the street used as scene is marked in white.

behavioral effects.

Therefore, we decided to accept less complete coverage of R4 by recording all sequences at one street section depicted in Figure 3, yielding two advantages: First, we reduce the complexity of the acquisition process. Second, we achieve maximum control over the sequence content.

Instead of sampling from the real world distribution, our recording strategy aims to include each difficult combination of adverse effects at least once. We address the assumed weak representativeness of real-world sampling by constructing a sample we assume to be more representative.

We assume the most critical regions for driving decisions are the road and sidewalks next to it. We approximate these effects at a 300 meters long street section with a T-junction at around 230 m. The scene contains small and tall buildings as well as trees causing complex shadows, influencing R1 (light) and R2 (weather). The turning situation at the junction allows to change the angle of the sun with respect to the car. We had to accept the weather occurring during the recording days scheduled at various times of the year relying on weather forecasts.

To address R1 and R3, we recorded on six days distributed over three seasons. The effects caused by time of year mainly included change of vegetation (leaves versus no leaves) and pedestrian clothing (more versus less). A minor role played the light intensity and direction of the sun which could be varied more strongly by time of day. Sequences where recorded on relatively regular intervals between sunrise and sunset during the day (R1). To simulate a situation of heavy rain directly followed by direct sun, we had firefighters emptying their tanks on the road to create very wet roads including large puddles.

To address R5, we hired around 40 actors consisting of infants, kids, teenagers as well as adults. Each of them were asked to attend a subset of our recording sessions. A main point was to include as much variance in looks as possible, roughly approximating the distribution of the real world. We added a variety of animals and props, including large and small dogs, toys, balls of various sizes, umbrellas, bags and exotic items such as a large mirror. To simulate actual car crashes with pedestrians, we used a moving puppet which could be overrun by a car without damaging the vehicle. Our selection of vehicles contained a skateboard, skates, kid's strollers, bikes, a motorbike, trucks and cars. We asked the respective drivers to arrange situations with approaching, queuing and overtaking vehicles and different turning and parking situations at two junctions.

To generate a large range of situations we varied the number of pedestrians and their behavior in each sequence. Therefore, we supplied the actors with a set of 10 instructions, each coming with a few examples: they had to choose a starting location, a speed, a speed change while they are seen by the cameras, a direction, a change of direction, an intention, a pose, a prop to act with, one or more co-actors and a scenario in which this combination seems likely to them. Each time a new sequence was recorded, the director went through these items and asked everyone to select a new combination.

A number of situations could not properly be addressed: High-speed driving on highways is not included as well as driving at deep night. We do not feature situations with bridges and tunnels which are of special interest due to sudden changes of lighting. Snow and fog were not available at any of our recording sessions. Further, complex turning situations with a varying number of junctions could not be addressed. Finally, there are most likely scenarios we did not think of at all mainly due to cultural biases and differences in technologies in different countries.

However, our dataset is the first that was specifically designed to sample from a relevant subset of urban autonomous driving situations. The limitations in the parameter space have been chosen carefully in order to make a systematic variation of the remaining parameters more feasible. All situations are well-motivated and represent a meaningful subset of sequences.

4. Results

In this Section, we verify that the recorded data meets our requirements (Section 4.1), compare our results with the four most relevant datasets (Section 4.2) and select a subset of our data for a benchmark to be published on a dedicated website (Section 4.3).

4.1. Dataset Overview

From around 200 sequences comprising 2.5 million image pairs at 200Hz, we selected 55 partial sequences with a total of 3563 image pairs at 25 Hz. Each sequence contains between 19 and 100 consecutive frames

We had clear skies, cloudy days and overcast days including occasional light rain. Including the simulated rain, we cover day and night, as well as dry and wet roads including combinations of all situations.

Although we recorded at several times of year, the light effects on images where negligible, mainly because the ex-

Figure 4. Sample images from our dataset. It comprises a large number of difficult light and weather situations such as low light, lens-flares, rain, and wet streets.

posure times could be adjusted to the amount of light available without requiring long exposure times causing motion blur. The different trajectories of the sun could as well be simulated by recording different angles of road as well as different times of day. A comparison of some weather situations is shown in Figure 4.

Depending on temperatures the actors dressed appropriately. Since we had some very hot and very cold recording sessions, the diversity of clothes is very high. Based on the results of the acting instructions we selected both very common as well as really surprising sequences for the benchmark. In the final selection, we have between 0 and 16 of these actors populating each sequence.

Since our dataset only allows for LIDAR-measured ground truth in the static parts of the sequence, we created manually annotated masks via crowdsourcing. We labeled each dynamic region with either *vehicle*, *person*, *other* and *unsure*. About 6% of all pixels contain dynamic motion. About 5% of all pixels are equally distributed between vehicle and person. Example labels are shown in Figure 2.

4.2. Dataset Comparison

With our dataset, we add difficult light and weather situations to the mix. Complementary to other datasets we focus on robustness given a high-end camera system: HFR at 200 Hz enables research in new methods focusing on both large and very small motions; HR challenges existing algorithms since large images often cause them to use prohibitive amounts of time and memory; HDR is a technology soon to become mainstream and should hence be a standard for new datasets. It reduces the challenges caused by strong light effects so that algorithms have a more realistic chance at delivering good results in adverse conditions. As a downside, we did not record color images because the required hardware was not commercially available at that time. Other limitations are the dynamic regions which contain only estimated ground truth as well as the restriction to a single location. The ground truth accuracy is limited due to a doubled image width compared to KITTI, containing regions with uncertainties of around 3 px.

Comparison to KITTI. KITTI uses a similar ground truth acquisition strategy to ours based on LIDAR. In contrast to our approach (cf. Section 3.2), KITTI scans continuously while driving with a car-mounted Velodyne device. The main advantage of their approach is that the extrinsic calibration with the LIDAR can be carried out once and the car can record ground truth without a prior scanning step. Yet, KITTI has a sparser and less accurate point cloud (± 2 cm according to the manufacturer) which is densified using a semi-automatic ICP step. The manual interaction in KITTI comes in with the cleaning and fine-tuning of the aggregation of the multiple scans for each frame. We only need to clean up the point cloud once, but need to manually establish 2D to 3D correspondences. Our approach delivers a very dense point cloud at a high accuracy of ± 1 cm. This allows for a higher image resolution because more 3D points fall into the 2D pixel locations.

KITTI further comes with a number of additional annotations such as semantic segmentations and scene flow (cf. e.g. [32]). To ensure representativeness, they used an unsupervised clustering approach applied to the recorded image sequences. In contrast, we chose to design the dataset by controlling the environment (cf. Section 3.3).

With respect to dataset content (R1-5), KITTI focuses on scenes with good weather and sufficient light, whereas we also address bad weather and low light. In KITTI, difficult light situations occur in saturated pixels and specular reflections e.g. on other vehicles. Dynamically moving parts are excluded from the ground truth and not treated further. We include pixel-accurate contours of the regions with coarse estimates of ground truth, allowing e.g. for analysis of foreground fattening metrics [16]. The scene locations in KITTI vary significantly, while the time of year does not change noticeably. Our dataset is complementary in that the location does not change, but time of year is covered well. Overfitting (R6) is reduced in KITTI by a relatively large amount of short sequences (around 800 in total) and separating the dataset into a training and a hidden test ground truth set. The resolution (R7, 1240×380), dynamic range (R8, 8 bit) and frame-rate (R9, 10 Hz) of the images are relatively low compared to ours. Uncertainties (R10) are not available in KITTI, but according to the authors, most disparities are around 3 px accurate with some flow vectors containing relative errors around 5% of the magnitude. Scaling their disparities up to our image width this would result in accu-

	Stereo						Labels	Flow				
	Middlebury		KITTI		MPI-Sintel	**Ours**	Cityscapes	Middlebury	KITTI		MPI-Sintel	**Ours**
Max. Resolution	1396x1110	2964x1988	1240x380	1240x380	1024x436	2560x1080	2048x1024	640x480	1240x380	1240x380	1024x436	2560x1080
Dynamic Range [bit]	14 to 8	14 to 8	8	8	8	16 to 8	16 to 8	12 to 8	8	8	8	16 to 8
Baseline [cm]	8-16	14-40	54	54	10	30	NA	NA	NA	NA	NA	NA
Max GT Length [px]	240	800	150	150	yet unkown	200	NA	35	150	250	400	12
Overall Accuracy [px]	1/4 (low res)	1/5	3	3	Optimal	Selectable	3 (contours)	1/10-1/60	3 (or 5%)	3 (or 5%)	Optimal	Selectable
Year of Publication	2003-6	2014	2012	2015	2016	**2016**	2015	2007	2012	2015	2012	**2016**
R1 (Radiometry)	+	+	+	+	+	++	+	-	+	+	+	++
R2 (Distractions)	-	-	+	+	+	++	-	-	+	+	+	++
R3 (Time of Year)			-	-		++	-		-	-		++
R4 (Location)			+	+		-	++		+	+		-
R5 (Geometry)	+	++	+	+	++	++	++	+	+	+	++	++
R6 (Overfitting)	-	++	+	+	++	++	++	+	+	+	++	++
R7 (Resolution)	+	++	-	-	+	++	++	-	-	-	+	++
R8 (Dynamic Range)	+	++	-	-	+	++	++	-	-	-	+	++
R9 (Frame Rate)						++			-	-	+	++
R10 (Accuracy)	+	++	-	-	++	+		+	-	-	++	+
R11 (Accessibility)	++	++	++	++	++	++	++	++	++	++	++	++
R12 (Comparison)	+	++	++	++	++	++		+	++	++	++	++

Figure 5. Overview of all dataset and benchmark requirements as defined in Section 3.1 and the most relevant benchmarks for autonomous driving. This Figure has been created in close collaboration with the respective paper authors. Green: best or equivalent to others. Yellow: good compared to others. Red: suboptimal properties compared to other entries. Gray: not applicable or unknown. Note that the MPI-Sintel stereo dataset is not yet published.

racies around 6 px. The benchmark website and dataset are highly accessible (R11) and allow for comparison based on a number of metrics (R12).

Comparison to MPI-Sintel. Although not perfectly realistic for autonomous driving, we include a comparison because MPI-Sintel encouraged a number of well-performing algorithms for large-displacement optical flow. Another important advantage of MPI-Sintel is that the ground truth can be parameterized and re-rendered arbitrarily (R6) with perfect ground truth (R10).

The current version comprises 35 sequences and already comes with two rendered passes including e.g. more and less realistic material properties (R1). Distractions (R2) can be found (fog, snow, etc.) but are of no special focus. Location and time of year play no meaningful role due to the artificial nature of the images. Resolution and dynamic range are as low as those of KITTI, while a slightly higher frame rate is available. Accessibility and comparison tools are comparable to those of KITTI.

Although not relevant for fixed-focus cameras in autonomous driving, one interesting aspect is that the focal length varies throughout the sequences including zooms, which are not present in other datasets.

Comparison to Middlebury v3. The most recent Middlebury stereo dataset uses high-end cameras and a very robust measurement system based on structured light scans of relatively large setups. The main advantage of this approach is its accuracy (R10) and versatility at least for medium-sized scenes. It further comes with multiple exposures and light settings (R1) for each dataset, supporting new research into learning better data terms [50]. Geometric complexity, resolution, dynamic range are all high. The downsides are that it does not come with flow ground truth, is of relatively small size (R6) and was recorded indoors with controlled lighting (R2-4). To reduce overfitting, uncertainties are supplied and each algorithm can only be submitted once to the test set. Accessibility and comparison tools are comparable to those of KITTI.

Comparison to Cityscapes. Finally, the Cityscapes Dataset is somewhat unrelated as it does not come with flow or stereo ground truth. On the other hand, it comes with pre-computed depth maps based on SGM [7] and covers scene labeling very thoroughly. This dataset is highly relevant for autonomous driving with respect to context awareness. Therefore, Cityscapes is very complementary to all other related datasets.

4.3. Benchmarking Approach

In this Section we discuss how we address requirements R11-13. To create a benchmark focusing on stereo and flow (cf. Section 4.3) we selected frames for low-level effects with light and geometry rather than for high-level semantics. We further selected sequences at various driving speeds ranging from around 0-70 kph.

The radiometric challenges (cf. Figure 4) in our selected sequences comprise: Raindrops in the scene and on the windshield (R2); The windshield wiper obstructing the view on the scene (R2); Saturated pixels caused by the sun, headlights of approaching vehicles and road reflections (R1); Mirroring reflections in puddles, on car surfaces and objects carried around by pedestrians (R1,R2); Very dark sequences after sunset (R1).

Geometric challenges are mainly created by traffic participants (R5); Complex occlusion patterns caused by pedestrians and vehicles; Geometrically highly detailed objects such as fences, street lamps, and other elongated objects such as a blind man's stick; Very large displacements and disparities; Small and independently moving objects such as balls; Complex deformations such as opening umbrellas, flapping blankets, flying hair as well as a skate-

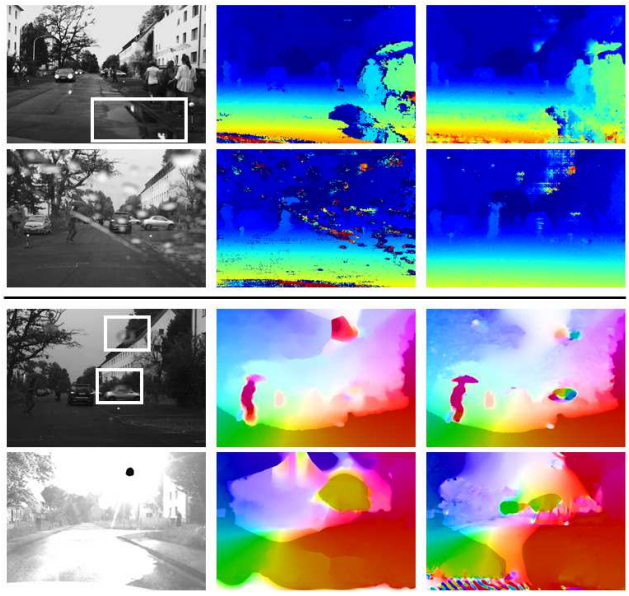

Figure 6. Example results for various stereo algorithms at difficult locations. The two top rows visualize stereo results of various algorithms; the two bottom rows show flow with the same color coding as in Figure 1. A closer look at the results can be found in the supplemental material.

boarder separating from his board; Standing, turning, parking and backwards-driving vehicles such as motorbikes, cars, vans and trucks; Suddenly opening vehicle doors; Large crowds of people running or walking at various distances on the road.

From all selected frame pairs we selected another subset of 10 key-frames for an example benchmarking based on four flow[4] and stereo algorithms[5] including most recent methods. As depicted in Figure 6 and detailed in the supplemental material, even most recent methods struggle with challenges such as raindrops on the windshield or intense glare.

For optimal accessibility (R11), we created a benchmarking website which allows for searching all sequences for relevant properties such as weather. We make available the full training dataset as well as uploaded results.

To enable comparison with other methods (R12), we implemented performance metrics currently available in KITTI and Middlebury. Additionally, for stereo we use a recently proposed set of semantically meaningful performance metrics such as edge fattening and surface smoothness [16] along with their respective visualizations.

To ease scientific dissemination of the results, researchers can upload their results and compare it to all publicly available datasets with ground truth as well as previously uploaded results. Especially the availability of results of existing methods allows for research in confidence analysis [13], post-processing [22] and aggregation [28]. Many visualizations can be downloaded in formats suitable for figures in publications.

Similar to other benchmarks, we avoid overfitting by only providing ground truth for half of all sequences. To create a variety of challenges, we reduced the frame-rate to 25Hz of another half of the public datasets. Finally, we removed the right frame from another half of the dataset to motivate research in strictly monoscopic algorithms.

Computing stereo or flow on all frames in the benchmark can be infeasible due to time and memory constraints for many algorithms. Hence, to improve accessibility and comparisons, for each sequence we selected representative frames as *challenge frames* which will be used to compute the rankings. Submission rules for the challenge frames will be aligned with the most recent Middlebury stereo challenge [39]. Both website and initial benchmark dataset including training and test images can be found at

http://hci-benchmark.org.

5. Conclusion

We designed and recorded a new stereo and flow dataset and extracted an initial benchmark subset comprising 28504 stereo pairs with stereo and flow ground truth with uncertainties for static regions. Dynamic regions, covering around 6% of all pixels, are manually masked out and annotated with approximate ground truth on 3500 pairs. Half of the ground truth is made available as training data. New stereo metrics and interactive results visualizations are accessible through our benchmark website. This way, we push the boundaries of what is currently achievable for large-scale outdoor stereo and flow reference data. This dataset is highly accurate compared to similar existing benchmarks. However, a small fraction (we estimate much fewer than 0.1% of all pixels) of our ground truth contain wrong values, mainly due to current technological limits in LIDAR. Future work will therefore focus on improving outlier detection tools and measurement setups. We are now working on more detailed labels including a full scene labeling and specialized pedestrian labels for action recognition as well as new metrics for temporal consistency. In the future, we will update our dataset with more sequences and more detailed masks for all kinds of radiometric as well as geometric challenges.

References

[1] S. Abraham and T. Hau. Towards autonomous high-precision calibration of digital cameras. In *Videometrics V, Proceedings of SPIE Annual Meeting*, volume 3174, pages 82–93. Citeseer, 1997. 4

[4] flow: Horn&Schunck[17], FlowFields [2], MDPFlow [47], Charb.[6]

[5] stereo: Elas[11], SPS-St[48], OCV-SGBM[15], MST [49]

[2] C. Bailer, B. Taetz, and D. Stricker. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *Computer Vision (ICCV). International Conference on Computer Vision (ICCV-15)*. IEEE, 2015. 8

[3] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, November 2010. 2

[4] J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994. 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, October 2012. 1, 2

[6] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *Image Processing, IEEE Transactions on*, 6(2):298–311, 1997. 8

[7] 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. 1, 2, 3, 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3

[9] A. Donath and D. Kondermann. Is crowdsourcing for optical flow ground truth generation feasible? In *Lecture Notes in Computer Science*, pages 193–202. Springer Science and Business Media, 2013. 2

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), June 2012. 1, 2

[11] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010. 8

[12] B. Gussefeld, D. Kondermann, C. Schwartz, and R. Klein. Are reflectance field renderings appropriate for optical flow evaluation? In *2014 IEEE International Conference on Image Processing (ICIP)*. Institute of Electrical & Electronics Engineers (IEEE), October 2014. 2

[13] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), June 2013. 8

[14] V. Haltakov, C. Unger, and S. Ilic. Framework for generation of synthetic ground truth data for driver assistance applications. In *GCPR*, September 2013. 2

[15] H. Hirschmueller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30:328–41, 2008. 8

[16] K. Honauer, L. Maier-Hein, and D. Kondermann. The HCI stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015. 3, 6, 8

[17] B. K. Horn and B. G. Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981. 8

[18] A. Humayun, O. Mac Aodha, and G. J. Brostow. Learning to find occlusion regions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2161–2168. IEEE, 2011. 2

[19] B. Jähne. *Digitale Bildverarbeitung*. Springer, Berlin, 7 edition, 2012. 4

[20] B. Jähne and M. Schwarzbauer. Noise equalisation and quasi loss-less image data compression–or how many bits needs an image sensor? *tm-Technisches Messen*, 83(1):16–24, 2016. 4

[21] N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011. 2

[22] C. Kondermann, D. Kondermann, and C. Garbe. Postprocessing of optical flows via surface measures and motion inpainting. In *Lecture Notes in Computer Science*, pages 355–364. Springer Science and Business Media, 2008. 8

[23] D. Kondermann, S. Abraham, G. Brostow, W. Förstner, S. Gehrig, A. Imiya, B. Jähne, F. Klose, M. Magnor, H. Mayer, R. Mester, T. Pajdla, R. Reulke, and H. Zimmer. On performance analysis of optical flow algorithms. In *Lecture Notes in Computer Science*, pages 329–355. Springer Science and Business Media, 2012. 2

[24] D. Kondermann, R. Nair, S. Meister, W. Mischler, B. Güssefeld, K. Honauer, S. Hofmann, C. Brenner, and B. Jähne. Stereo ground truth with error bars. In *Computer Vision – ACCV 2014*, pages 595–610. Springer Science and Business Media, 2015. 2, 3, 4

[25] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. 2

[26] P. Leclercq and J. Morris. Assessing stereo algorithm accuracy. *Proceedings of Image and Vision Computing*, 2:89–93, 2002. 2

[27] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss. Human-assisted motion annotation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), June 2008. 2

[28] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow. Learning a confidence measure for optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (early access articles)*, PP, 2012. 8

[29] B. McCane, K. Novins, D. Crannitch, and B. Galvin. On benchmarking optical flow. http://of-eval.sourceforge.net/, 2001. 2

[30] S. Meister, S. Izadi, P. Kohli, M. Hämmerle, C. Rother, and D. Kondermann. When can we use KinectFusion for ground truth acquisition? In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, Workshops & Tutorials*, 2012. 2

[31] S. Meister and D. Kondermann. Real versus realistically rendered scenes for optical flow evaluation. In *Electronic Media Technology (CEMT), 2011 14th ITG Conference on*, 2011. 2

[32] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[33] S. Morales and R. Klette. A third eye for performance evaluation in stereo sequence analysis. In *Computer Analysis of Images and Patterns*, pages 1078–1086. Springer, 2009. 2

[34] 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. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[35] N. Onkarappa and A. D. Sappa. Synthetic sequences and ground-truth flow field generation for algorithm validation. *Multimedia Tools and Applications*, pages 1–15, 2013. 2

[36] M. Otte and H.-H. Nagel. Optical flow estimation: advances and comparisons. In *Computer Vision—ECCV'94*, pages 49–60. Springer, 1994. 2

[37] D. Pfeiffer and U. Franke. Towards a global optimal multi-layer stixel representation of dense 3d data. In *BMVC*, pages 1–12, 2011. 2

[38] D. Pfeiffer, S. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 297–304. IEEE, 2013. 2

[39] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer, 2014. 1, 2, 8

[40] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–195. IEEE, 2003. 2

[41] A. Sellent, D. Kondermann, S. Simon, S. Baker, G. Dedeoglu, O. Erdler, P. Parsonage, C. Unger, and W. Niehsen. Optical flow estimation versus motion estimation. 2012. 3

[42] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[43] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behavior on synthetic and real-world stereo sequences. In *23rd International Conference of Image and Vision Computing New Zealand (IVCNZ '08)*, pages 1–6, 2008. 2

[44] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In *2008 23rd International Conference Image and Vision Computing New Zealand*. Institute of Electrical & Electronics Engineers (IEEE), November 2008. 2

[45] W. Wachenfeld and H. Winner. The release of autonomous vehicles, 2016. to be published. 3

[46] J. Wulff, D. J. Butler, G. B. Stanley, and M. J. Black. Lessons and insights from creating a synthetic optical flow benchmark. In A. Fusiello et al. (Eds.), editor, *ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*, Part II, LNCS 7584, pages 168–177. Springer-Verlag, October 2012. 2

[47] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1744–1757, 2012. 8

[48] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Computer Vision–ECCV 2014*, pages 756–771. Springer, 2014. 8

[49] Q. Yang. A non-local cost aggregation method for stereo matching. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1402–1409. IEEE, 2012. 8

[50] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *CoRR*, abs/1510.05970, 2015. 7