Are we ready for Autonomous Driving?
The KITTI Vision Benchmark Suite

Andreas Geiger and Philip Lenz
Karlsruhe Institute of Technology
{geiger,lenz}@kit.edu

Raquel Urtasun
Toyota Technological Institute at Chicago
rurtasun@ttic.edu

Abstract

Today, visual recognition systems are still rarely employed in robotics applications. Perhaps one of the main reasons for this is the lack of demanding benchmarks that mimic such scenarios. In this paper, we take advantage of our autonomous driving platform to develop novel challenging benchmarks for the tasks of stereo, optical flow, visual odometry / SLAM and 3D object detection. Our recording platform is equipped with four high resolution video cameras, a Velodyne laser scanner and a state-of-the-art localization system. Our benchmarks comprise 389 stereo and optical flow image pairs, stereo visual odometry sequences of 39.2 km length, and more than 200k 3D object annotations captured in cluttered scenarios (up to 15 cars and 30 pedestrians are visible per image). Results from state-of-the-art algorithms reveal that methods ranking high on established datasets such as Middlebury perform below average when being moved outside the laboratory to the real world. Our goal is to reduce this bias by providing challenging benchmarks with novel difficulties to the computer vision community. Our benchmarks are available online at: **www.cvlibs.net/datasets/kitti**

1. Introduction

Developing autonomous systems that are able to assist humans in everyday tasks is one of the grand challenges in modern computer science. One example are autonomous driving systems which can help decrease fatalities caused by traffic accidents. While a variety of novel sensors have been used in the past few years for tasks such as recognition, navigation and manipulation of objects, visual sensors are rarely exploited in robotics applications: Autonomous driving systems rely mostly on GPS, laser range finders, radar as well as very accurate maps of the environment.

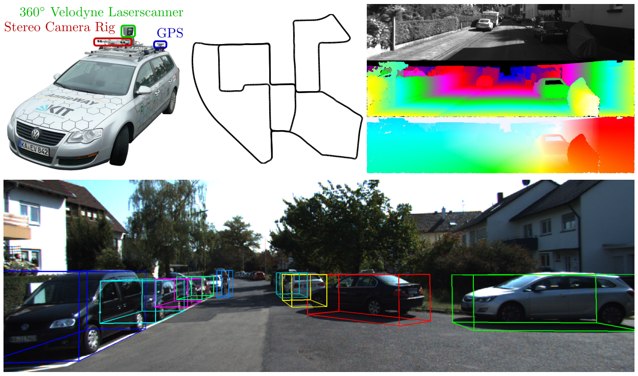


Figure 1. **Recording platform** with sensors (top-left), **trajectory** from our visual odometry benchmark (top-center), **disparity** and **optical flow** map (top-right) and **3D object** labels (bottom).

In the past few years an increasing number of benchmarks have been developed to push forward the performance of visual recognitions systems, e.g., Caltech-101 [17], Middlebury for stereo [41] and optical flow [2] evaluation. However, most of these datasets are simplistic, e.g., are taken in a controlled environment. A notable exception is the PASCAL VOC challenge [16] for detection and segmentation.

In this paper, we take advantage of our autonomous driving platform to develop novel challenging benchmarks for stereo, optical flow, visual odometry / SLAM and 3D object detection. Our benchmarks are captured by driving around a mid-size city, in rural areas and on highways. Our recording platform is equipped with two high resolution stereo camera systems (grayscale and color), a Velodyne HDL-64E laser scanner that produces more than one million 3D points per second and a state-of-the-art OXTS RT 3003 localization system which combines GPS, GLONASS, an IMU and RTK correction signals. The cameras, laser scanner and localization system are calibrated and synchronized, providing us with accurate ground truth. Table 1 summarizes our benchmarks and provides a comparison to existing datasets.

Our **stereo** matching and **optical flow** estimation benchmark comprises 194 training and 195 test image pairs at a resolution of 1240×376 pixels after rectification with semi-dense (50%) ground truth. Compared to previous datasets [41, 2, 30, 29], this is the first one with realistic non-synthetic imagery and accurate ground truth. Difficulties include non-lambertian surfaces (e.g., reflectance, transparency) large displacements (e.g., high speed), a large variety of materials (e.g., matte vs. shiny), as well as different lighting conditions (e.g., sunny vs. cloudy).

Our **3D visual odometry / SLAM** dataset consists of 22 stereo sequences, with a total length of 39.2 km. To date, datasets falling into this category are either monocular and short [43] or consist of low quality imagery [42, 4, 35]. They typically do not provide an evaluation metric, and as a consequence there is no consensus on which benchmark should be used to evaluate visual odometry / SLAM approaches. Thus often only qualitative results are presented, with the notable exception of laser-based SLAM [28]. We believe a fair comparison is possible in our benchmark due to its large scale nature as well as the novel metrics we propose, which capture different sources of error by evaluating error statistics over all sub-sequences of a given trajectory length or driving speed.

Our **3D object** benchmark focuses on computer vision algorithms for object detection and 3D orientation estimation. While existing benchmarks for those tasks do not provide accurate 3D information [17, 39, 15, 16] or lack realism [33, 31, 34], our dataset provides accurate 3D bounding boxes for object classes such as cars, vans, trucks, pedestrians, cyclists and trams. We obtain this information by manually labeling objects in 3D point clouds produced by our Velodyne system, and projecting them back into the image. This results in tracklets with accurate 3D poses, which can be used to asses the performance of algorithms for 3D orientation estimation and 3D tracking.

In our **experiments**, we evaluate a representative set of state-of-the-art systems using our benchmarks and novel metrics. Perhaps not surprisingly, many algorithms that do well on established datasets such as Middlebury [41, 2] struggle on our benchmark. We conjecture that this might be due to their assumptions which are violated in our scenarios, as well as overfitting to a small set of training (test) images.

In addition to the benchmarks, we provide MATLAB/C++ development kits for easy access. We also maintain an up-to-date online evaluation server[1]. We hope that our efforts will help increase the impact that visual recognition systems have in robotics applications.

2. Challenges and Methodology

Generating large-scale and realistic evaluation benchmarks for the aforementioned tasks poses a number of challenges, including the collection of large amounts of data in real time, the calibration of diverse sensors working at different rates, the generation of ground truth minimizing the amount of supervision required, the selection of the appropriate sequences and frames for each benchmark as well as the development of metrics for each task. In this section we discuss how we tackle these challenges.

2.1. Sensors and Data Acquisition

We equipped a standard station wagon with two color and two grayscale PointGrey Flea2 video cameras (10 Hz, resolution: 1392×512 pixels, opening: $90^\circ \times 35^\circ$), a Velodyne HDL-64E 3D laser scanner (10 Hz, 64 laser beams, range: 100 m), a GPS/IMU localization unit with RTK correction signals (open sky localization errors < 5 cm) and a powerful computer running a real-time database [22].

We mounted all our cameras (i.e., two units, each composed of a color and a grayscale camera) on top of our vehicle. We placed one unit on the left side of the rack, and the other on the right side. Our camera setup is chosen such that we obtain a baseline of roughly 54 cm between the same type of cameras and that the distance between color and grayscale cameras is minimized (6 cm). We believe this is a good setup since color images are very useful for tasks such as segmentation and object detection, but provide lower contrast and sensitivity compared to their grayscale counterparts, which is of key importance in stereo matching and optical flow estimation.

We use a Velodyne HDL-64E unit, as it is one of the few sensors available that can provide accurate 3D information from moving platforms. In contrast, structured-light systems such as the Microsoft Kinect do not work in outdoor scenarios and have a very limited sensing range. To compensate egomotion in the 3D laser measurements, we use the position information from our GPS/IMU system.

2.2. Sensor Calibration

Accurate sensor calibration is key for obtaining reliable ground truth. Our calibration pipeline proceeds as follows: First, we calibrate the four video cameras intrinsically and extrinsically and rectify the input images. We then find the 3D rigid motion parameters which relate the coordinate system of the laser scanner, the localization unit and the reference camera. While our Camera-to-Camera and GPS/IMU-to-Velodyne registration methods are fully automatic, the Velodyne-to-Camera calibration requires the user to manually select a small number of correspondences between the laser and the camera images. This was necessary as existing techniques for this task are not accurate enough to compute ground truth estimates.

Camera-to-Camera calibration. To automatically calibrate the intrinsic and extrinsic parameters of the cameras, we mounted checkerboard patterns onto the walls of our garage and detect corners in our calibration images. Based on gradient information and discrete energy-minimization, we assign corners to checkerboards, match them between

[1] www.cvlibs.net/datasets/kitti

Stereo Matching	type	#images	resolution	ground truth	uncorrelated	metric
EISATS [30]	synthetic	498	0.3 Mpx	dense		
Middlebury [41]	laboratory	38	0.2 Mpx	dense	✓	✓
Make3D Stereo [40]	real	257	0.8 Mpx	0.5 %	✓	✓
Ladicky [29]	real	70	0.1 Mpx	manual	✓	
Proposed Dataset	**real**	**389**	**0.5 Mpx**	**50 %**	✓	✓

Optical Flow	type	#images	resolution	ground truth	uncorrelated	metric
EISATS [30]	synthetic	498	0.3 Mpx	dense		
Middlebury [2]	laboratory	24	0.2 Mpx	dense	✓	✓
Proposed Dataset	**real**	**389**	**0.5 Mpx**	**50 %**	✓	✓

Visual Odometry / SLAM	setting	#sequences	length	#frames	resolution	ground truth	metric
TUM RGB-D [43]	indoor	27	0.4 km	65k	0.3 Mpx	✓	✓
New College [42]	outdoor	1	2.2 km	51k	0.2 Mpx		
Malaga 2009 [4]	outdoor	6	6.4 km	38k	0.8 Mpx	✓	
Ford Campus [35]	outdoor	2	5.1 km	7k	1.0 Mpx	✓	
Proposed Dataset	**outdoor**	**22**	**39.2 km**	**41k**	**0.5 Mpx**	✓	✓

Object Detection / 3D Estimation	#categories	avg. #labels/category	occlusion labels	3D labels	orientations
Caltech 101 [17]	101	40-800			
MIT StreetScenes [3]	9	3,000			
LabelMe [39]	3997	60			
ETHZ Pedestrian [15]	1	12,000			
PASCAL 2011 [16]	20	1,150	✓		
Daimler [8]	1	56,000	✓		
Caltech Pedestrian [13]	1	350,000	✓		
COIL-100 [33]	100	72		✓	72 bins
EPFL Multi-View Car [34]	20	90		✓	90 bins
Caltech 3D Objects [31]	100	144		✓	144 bins
Proposed Dataset	**2**	**80,000**	✓	✓	**continuous**

Table 1. **Comparison of current State-of-the-Art Benchmarks and Datasets.**

the cameras and optimize all parameters by minimizing the average reprojection error [19].

Velodyne-to-Camera calibration. Registering the laser scanner with the cameras is non-trivial as correspondences are hard to establish due to the large amount of noise in the reflectance values. Therefore we rely on a semi-automatic technique: First, we register both sensors using the fully automatic method of [19]. Next, we minimize the number of disparity outliers with respect to the top performing methods in our benchmark jointly with the reprojection errors of a few manually selected correspondences between the laser point cloud and the images. As correspondences, we select edges which can be easily located by humans in both domains (i.e., images and point clouds). Optimization is carried out by drawing samples using Metropolis-Hastings and selecting the solution with the lowest energy.

GPS/IMU-to-Velodyne calibration. Our GPS/IMU to Velodyne registration process is fully automatic. We cannot rely on visual correspondences, however, if motion estimates from both sensors are provided, the problem becomes identical to the well-known *hand-eye calibration problem*, which has been extensively explored in the robotics community [14]. Making use of ICP, we accurately register laser point clouds of a parking sequence, as this provides a large variety of orientations and translations necessary to well condition the minimization problem. Next, we randomly sample 1000 pairs of poses from this sequence and obtain the desired result using [14].

2.3. Ground Truth

Having calibrated and registered all sensors, we are ready to generate ground truth for the individual benchmarks shown in Fig. 1.

To obtain a high **stereo** and **optical flow** ground truth density, we register a set of consecutive frames (5 before and 5 after the frame of interest) using ICP. We project the accumulated point clouds onto the image and automatically remove points falling outside the image. We then manually remove all ambiguous image regions such as windows and fences. Given the camera calibration, the corresponding disparity maps are readily computed. Optical flow fields are obtained by projecting the 3D points into the next frame. For both tasks we evaluate both non-occluded pixels as well as all pixels for which ground truth is available. Our non-occluded evaluation excludes all surface points falling outside the image plane. Points occluded by objects within the same image could not be reliably estimated in a fully automatic manner due to the properties of the laser scanner. To avoid artificial errors, we do not interpolate the ground truth disparity maps and optical flow fields, leading to a $\sim 50\%$ average ground truth density.

The ground truth for **visual odometry/SLAM** is directly given by the output of the GPS/IMU localization unit pro-

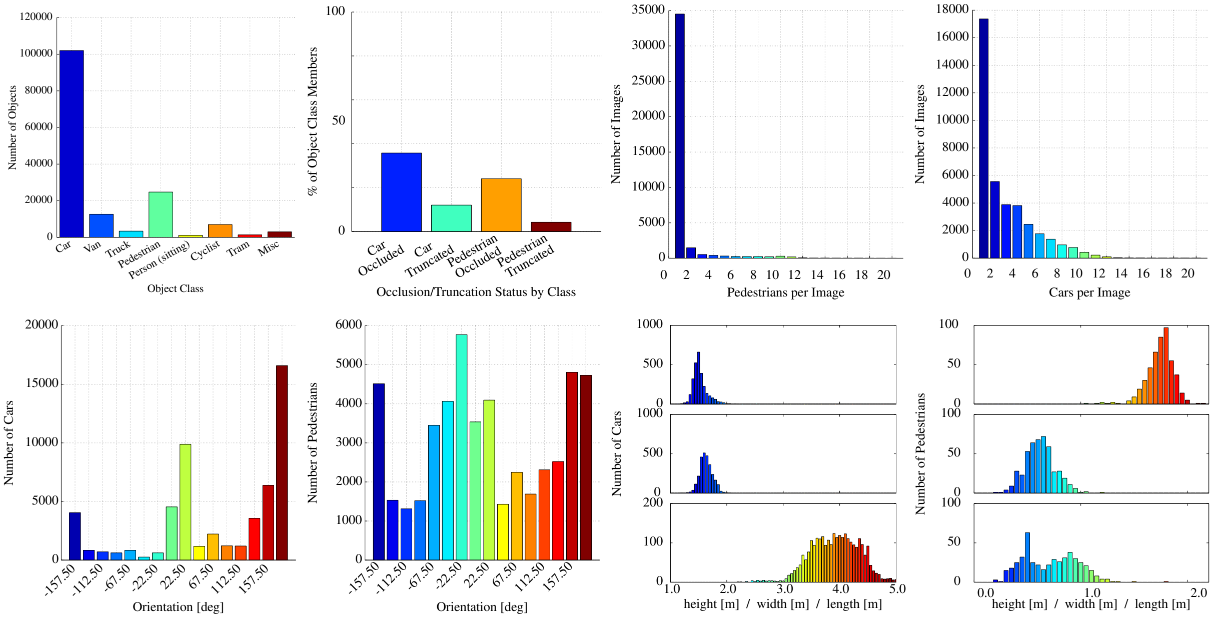


Figure 2. **Object Occurence and Object Geometry Statistics of our Dataset.** This figure shows (from left to right and top to bottom): The different types of objects occuring in our sequences, the power-law shaped distribution of the number of instances within an image and the orientation histograms and object size distributions for the two most predominant categories 'cars' and 'pedestrians'.

jected into the coordinate system of the left camera after rectification.

To generate **3D object** ground-truth we hired a set of annotators, and asked them to assign tracklets in the form of 3D bounding boxes to objects such as cars, vans, trucks, trams, pedestrians and cyclists. Unlike most existing benchmarks, we do not rely on online crowd-sourcing to perform the labeling. Towards this goal, we create a special purpose labeling tool, which displays 3D laser points as well as the camera images to increase the quality of the annotations. Following [16], we asked the annotators to additionally mark each bounding box as either visible, semi-occluded, fully occluded or truncated. Statistics of our labeling effort are shown in Fig. 2.

2.4. Benchmark Selection

We collected a total of $\sim$ 3 TB of data from which we select a representative subset to evaluate each task. In our experiments we currently concentrate on grayscale images, as they provide higher quality than their color counterparts.

For our **stereo** and **optical flow** benchmarks we select a subset of the sequences where the environment is static. To maximize diversity, we perform k-means ($k = 400$) clustering on the data using a novel representation, and chose the elements closest to the center of each cluster for the benchmark. We describe each image using a 144-dimensional image descriptor, obtained by subdividing the image into 12×4 rectangular blocks and computing the average disparity and optical flow displacement for each block. After removing scenes with bad illumination conditions as, e.g., tunnels, we obtain 194 training and 195 test image pairs for both benchmarks.

For our **visual odometry / SLAM** evaluation we select long sequences of varying speed with high-quality localization, yielding a set of 41.000 frames captured at 10 fps and a total driving distance of 39.2 km with frequent loop closures which are of interest in SLAM.

Our **3D object detection and orientation** estimation benchmark is chosen according to the number of non-occluded objects in the scene, as well as the entropy of the object orientation distribution. High entropy is desirable in order to ensure diversity. Towards this goal we utilize a greedy algorithm: We initialize our dataset $\mathcal{X}$ to the empty set $\emptyset$ and iteratively add images using the following rule

$$\mathcal{X} \leftarrow \mathcal{X} \cup \operatorname*{argmax}_{x} \left[\alpha \cdot noc(x) + \frac{1}{C} \sum_{c=1}^{C} H_c\left(\mathcal{X} \cup x\right) \right] \qquad (1)$$

where $\mathcal{X}$ is the current set, x is an image from our dataset, $noc(x)$ stands for the number of non-occluded objects in image x and C denotes the number of object classes. H_c is the entropy of class c with respect to orientation (we use $8/16$ orientation bins for pedestrians/cars). We further ensure that images from one sequence do not appear in both training and test set.

2.5. Evaluation Metrics

We evaluate state-of-the-art approaches utilizing a diverse set of metrics. Following [41, 2] we evaluate **stereo** and **optical flow** using the average number of erroneous pixels in terms of disparity and end-point error. In contrast to [41, 2], our images are not downsampled. Therefore, we employ a disparity/end-point error threshold of $\tau \in \{2,..,5\}$ px for our benchmark, with $\tau = 3$ px the default setting which takes into consideration almost all calibration and laser measurement errors. We report errors for both non-occluded pixels as well as all pixels where ground-truth is available.

Evaluating **visual odometry/SLAM** approaches based on the error of the trajectory end-point can be misleading, as this measure depends strongly on the point in time where the error has been made, e.g., rotational errors earlier in the sequence lead to larger end-point errors. Kümmerle at al. [28] proposed to compute the average of all relative relations at a fixed distance. Here we extend this metric in two ways. Instead of combining rotation and translation errors into a single measure, we treat them separately. Furthermore, we evaluate errors as a function of the trajectory length and velocity. This allows for deeper insights into the qualities and failure modes of individual methods. Formally, our error metrics are defined as

$$E_{rot}(\mathcal{F}) = \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} \angle[(\hat{\mathbf{p}}_j \ominus \hat{\mathbf{p}}_i) \ominus (\mathbf{p}_j \ominus \mathbf{p}_i)] \quad (2)$$

$$E_{trans}(\mathcal{F}) = \frac{1}{|\mathcal{F}|} \sum_{(i,j)\in\mathcal{F}} \|(\hat{\mathbf{p}}_j \ominus \hat{\mathbf{p}}_i) \ominus (\mathbf{p}_j \ominus \mathbf{p}_i)\|_2 \quad (3)$$

where $\mathcal{F}$ is a set of frames (i, j), $\hat{\mathbf{p}} \in SE(3)$ and $\mathbf{p} \in SE(3)$ are estimated and true camera poses respectively, $\ominus$ denotes the inverse compositional operator [28] and $\angle[\cdot]$ is the rotation angle.

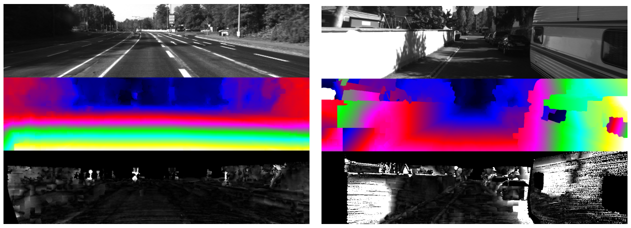
Our **3D object detection** and **orientation** estimation benchmark is split into three parts: First, we evaluate classical 2D object detection by measuring performance using the well established average precision (AP) metric as described in [16]. Detections are iteratively assigned to ground truth labels starting with the largest overlap, measured by bounding box intersection over union. We require true positives to overlap by more than 50% and count multiple detections of the same object as false positives. We assess the performance of jointly detecting objects and estimating their 3D orientation using a novel measure which we called the *average orientation similarity (AOS)*, which we define as:

$$AOS = \frac{1}{11} \sum_{r\in\{0,0.1,..,1\}} \max_{\tilde{r}:\tilde{r}\geq r} s(\tilde{r}) \quad (4)$$

Here, $r = \frac{TP}{TP+FN}$ is the PASCAL object detection recall, where detected 2D bounding boxes are correct if they overlap by at least 50% with a ground truth bounding box. The orientation similarity $s \in [0, 1]$ at recall r is a normalized ([0..1]) variant of the cosine similarity defined as

$$s(r) = \frac{1}{|\mathcal{D}(r)|} \sum_{i\in\mathcal{D}(r)} \frac{1 + \cos\Delta_\theta^{(i)}}{2} \delta_i \quad (5)$$

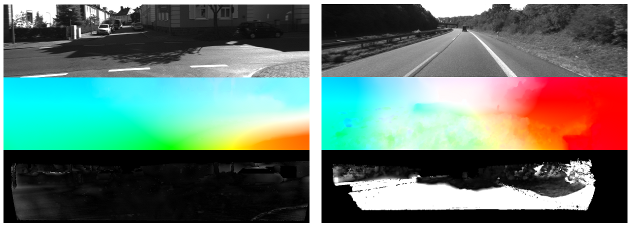
where $\mathcal{D}(r)$ denotes the set of all object detections at recall rate r and $\Delta_\theta^{(i)}$ is the difference in angle between estimated and ground truth orientation of detection i. To penalize multiple detections which explain a single object, we set $\delta_i = 1$ if detection i has been assigned to a ground truth bounding box (overlaps by at least 50%) and $\delta_i = 0$ if it has not been assigned.



(a) Best: < 1% errors

(b) Worst: 21% errors

Figure 3. **Stereo Results for PCBP [46].** Input image (top), estimated disparity map (middle), disparity errors (bottom). Error range: 0 px (black) to ≥ 5 px (white).



(a) Best: < 1% errors

(b) Worst: 59% errors

Figure 4. **Optical Flow Results for TGV2CENSUS [45].** Input image (top), estimated flow field (middle), end point errors (bottom). Error range: 0 px (black) to ≥ 5 px (white).

Finally, we also evaluate pure classification (16 bins for cars) and regression (continuous orientation) performance on the task of 3D object orientation estimation in terms of orientation similarity.

3. Experimental Evaluation

We run a representative set of state-of-the-art algorithms for each task. Interestingly, we found that algorithms ranking high on existing benchmarks often fail when confronted with more realistic scenarios. This section tells their story.

3.1. Stereo Matching

For stereo matching, we run global [26, 37, 46], semi-global [23], local [5, 20, 38] and seed-growing [27, 10, 9] methods. The parameter settings we have employed can be found on www.cvlibs.net/datasets/kitti. Missing disparities are filled-in for each algorithm using background interpolation [23] to produce dense disparity maps which can then be compared. As Table 2 shows, errors on our benchmark are higher than those reported on Middlebury [41], indicating

Stereo	Non-Occluded	All	Density
PCBP [46]	**4.72** %	**6.16** %	100.00 %
ITGV [37]	6.31 %	7.40 %	100.00 %
OCV-SGBM [5]	7.64 %	9.13 %	86.50 %
ELAS [20]	8.24 %	9.95 %	94.55 %
SDM [27]	10.98 %	12.19 %	63.58 %
GCSF [9]	12.06 %	13.26 %	60.77 %
GCS [10]	13.37 %	14.54 %	51.06 %
CostFilter [38]	19.96 %	21.05 %	100.00 %
OCV-BM [5]	25.39 %	26.72 %	55.84 %
GC+occ [26]	33.50 %	34.74 %	87.57 %

Optical Flow	Non-Occluded	All	Density
TGV2CENSUS [45]	**11.14** %	**18.42** %	100.00 %
HS [44]	19.92 %	28.86 %	100.00 %
LDOF [7]	21.86 %	31.31 %	100.00 %
C+NL [44]	24.64 %	33.35 %	100.00 %
DB-TV-L1 [48]	30.75 %	39.13 %	100.00 %
GCSF [9]	33.23 %	41.74 %	48.27 %
HAOF [6]	35.76 %	43.36 %	100.00 %
OCV-BM [5]	63.46 %	68.16 %	100.00 %
Pyramid-LK [47]	65.74 %	70.09 %	99.90 %

Table 2. **Stereo (left) and Optical Flow (right) Ranking from April 2, 2012.** Numbers denote the percentage of pixels with disparity error or optical flow end-point error (euclidean distance) larger than $\tau = 3px$, averaged over all test images. Here, *non-occluded* refers to pixels which remain inside the image after projection in both images and *all* denotes all pixels for which ground truth information is available. Density refers to the number of estimated pixels. Invalid disparities and flow vectors have been interpolated for comparability.

the increased level of difficulty of our real-world dataset. Interestingly, methods ranking high on Middlebury, perform particularly bad on our dataset, e.g., guided cost-volume filtering [38], pixel-wise graph cuts [26]. This is mainly due to the differences in the data sets: Since the Middlebury benchmark is largely well textured and provides a smaller label set, methods concentrating on accurate object boundary segmentation peform well. In contrast, our data requires more global reasoning about areas with little, ambiguous or no texture where segmentation performance is less critical. Purely local methods [5, 38] fail if fronto-parallel surfaces are assumed, as this assumption is often strongly violated in real-world scenes (e.g., road or buildings).

Fig. 3 shows the best and worst test results for the (currently) top ranked stereo method PCBP [46]. While small errors are made in natural environments due to the large degree of textureness, inner-city scenarios prove to be challenging. Here, the predominant error sources are image saturation (wall on the left), disparity shadows (RV occludes road) and non-lambertian surfaces (reflections on RV body).

3.2. Optical Flow Estimation

For optical flow we evaluate state-of-the-art variational [24, 6, 48, 44, 7, 9, 45] and local [5, 47] methods. The results of our experiments are summarized in Table 2. We observed that classical variational approaches [24, 44, 45] work best on our images. However, the top performing approach TGV2CENSUS [45] still produces about 11% of errors on average. As highlighted in Fig. 4, most errors are made in regions which undergo large displacements between frames, e.g., close range pixels on the street. Furthermore, pyramidal implementations lack the ability to estimate flow fields at higher levels of the pyramid due to missing texture. While best results are obtained at small motions (Fig. 4 left, flow $\leq$ 55 px), driving at high speed (Fig. 4 right, flow $\leq$ 176 px) leads to large displacements, which can not be reliably handled by any of the evaluated methods. We believe that to overcome these problems we need more complex models that utilize prior knowledge of the world. Previously hampered by the lack of sufficient training data, such approaches will become feasible in the near future with larger training sets as the one we provide.

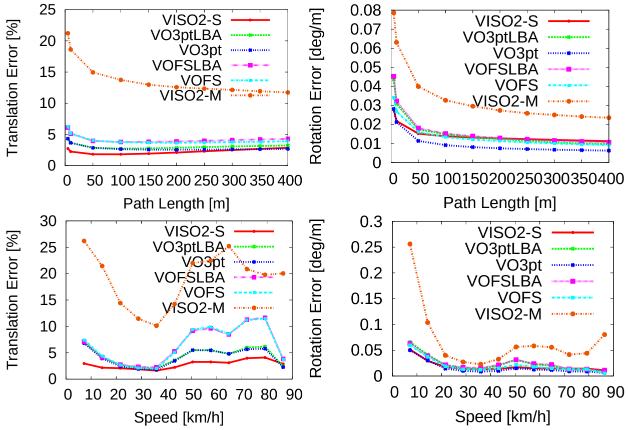
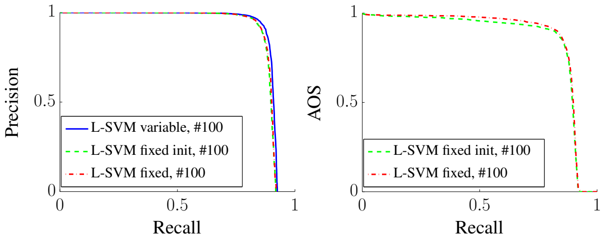


Figure 5. **Visual Odometry Evaluation.** Translation and rotation errors, averaged over all sub-sequences of a given length or speed.

3.3. Visual Odometry/SLAM

We evaluate five different approaches on our visual odometry / SLAM dataset: VISO2-S/M [21], a real-time stereo/monocular visual odometry library based on incremental motion estimates, the approach of [1] with and without Local Bundle Adjustment (LBA) [32] as well as the flow separation approach of [25]. All algorithms are comparable as none of them uses loop-closure information. All approaches use stereo with the exception of VISO2-M [21] which employs only monocular images. Fig. 5 depicts the rotational and translational errors as a function of the trajectory length and driving speed.

In our evaluation, VISO2-S [21] comes closest to the ground truth trajectories with an average translation error of 2.2% and an average rotation error of 0.016 deg/m. Akin to our optical flow experiments, large motion impacts performance, especially in terms of translation. With a recording



(a) Precision-Recall (b) Average Orientation Similarity

Figure 6. **Object Detection and Orientation Estimation Results.** Details about the metrics can be found in Sec. 2.5

rate of 10 frames per second, the vehicle moved up to 2.8 meters per frame. Additionally, large motions mainly occur on highways which are less rich in terms of 3D structure. Large errors at lower speeds stem from the fact that incremental or sliding-window based methods slowly drift over time, with the strongest *relative* impact at slow speeds. This problem can be easily alleviated if larger timespans are optimized when the vehicle moves slowly or is standing still. In our experiments, no ground truth information has been used to train the model parameters. We expect detecting loop closures, utilizing more enhanced bundle adjustment techniques as well as utilizing the training data for parameter fitting to further boost performance.

3.4. 3D Object Detection / Orientation Estimation

We evaluate object detection as well as joint detection and orientation estimation using average precision and average orientation similarity as described in Sec. 2.5. Our benchmark extracted from the full dataset comprises $12,000$ images with $40,000$ objects. We first subdivide the training set into 16 orientation classes and use 100 non-occluded examples per class for training the part-based object detector of [18] using three different settings: We train the model in an unsupervised fashion (*variable*), by initializing the components to the 16 classes but letting the components vary during optimization (*fixed init*) and by initializing the components and additionally fixing the latent variables to the 16 classes (*fixed*).

We evaluate all non- and weakly-occluded ($< 20\%$) objects which are neither truncated nor smaller than 40 px in height. We do not count detecting truncated or occluded objects as false positives. For our object detection experiment, we require a bounding box overlap of at least 50%, results are shown in Fig. 6(a). For detection and orientation estimation we require the same overlap and plot the average orientation similarity (Eq. 5) over recall for the two unsupervised variants (Fig. 6(b)). Note that the precision is an upper bound to the average orientation similarity.

Classification	Similarity	**Regression**	Similarity
SVM[11]	**0.93**	GP [36]	**0.92**
NN	0.85	SVM[11]	0.91
		NN	0.86

Table 3. **Object Orientation Errors for Cars.** Performance measured in terms of orientation similarity (Eq. 5). Higher is better.

Overall, we could not find any substantial difference between the part-based detector variants we investigated. All of them achieve high precision, while the recall seems to be limited by some hard to detect objects. We plan to extend our online evaluation to more complex scenarios such as semi-occluded or truncated objects and other object classes like vans, trucks, pedestrians and cyclists.

Finally, we also evaluate object orientation estimation. We extract 100 car instances per orientation bin, using 16 orientation bins. We compute HOG features [12] on all cropped and resized bounding boxes with 19×13 blocks, 8×8 pixel cells and 12 orientation bins. We evaluate multiple classification and regression algorithms and report average orientation similarity (Eq. 5). Table 3 shows our results. We found that for the classification task SVMs [11] clearly outperform nearest neighbor classification. For the regression task, Gaussian Process regression [36] performs best.

4. Conclusion and Future Work

Throwing new light on existing methods, we hope that the proposed benchmarks will complement others and help to reduce overfitting to datasets with little training or test examples and contribute to the development of algorithms that work well in practice. As our recorded data provides more information than compiled into the benchmarks so far, our intention is to gradually increase their difficulties. Furthermore, we also plan to include visual SLAM with loop-closure capabilities, object tracking, segmentation, structure-from-motion and 3D scene understanding into our evaluation framework.

References

[1] P. Alcantarilla, L. Bergasa, and F. Dellaert. Visual odometry priors for robust EKF-SLAM. In *ICRA*, 2010. 6

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92:1–31, 2011. 1, 2, 3, 4, 5

[3] S. M. Bileschi. Streetscenes: Towards scene understanding in still images. Technical report, MIT, 2006. 3

[4] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez. A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Auton. Robots*, 27:327–351, 2009. 2, 3

[5] G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 5, 6

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004. 6

[7] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *PAMI*, 33:500–513, March 2011. 6

[8] M. E. C. G. Keller and D. M. Gavrila. A new benchmark for stereo-based pedestrian detection. In *IV*, 2011. 3

[9] J. Cech, J. Sanchez-Riera, and R. P. Horaud. Scene flow estimation by growing correspondence seeds. In *CVPR*, 2011. 5, 6

[10] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *BenCOS*, 2007. 5, 6

[11] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Technical report, 2001. 7

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 7

[13] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. In *PAMI*, volume 99, 2011. 3

[14] F. Dornaika and R. Horaud. Simultaneous robot-world and hand-eye calibration. *Rob. and Aut.*, 1998. 3

[15] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007. 2, 3

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. 1, 2, 3, 4, 5

[17] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004. 1, 2, 3

[18] P. Felzenszwalb, R.Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained partbased models. *PAMI*, 32:1627–1645, 2010. 7

[19] A. Geiger, F. Moosmann, O. Car, and B. Schuster. A toolbox for automatic calibration of range and camera sensors using a single shot. In *ICRA*, 2012. 3

[20] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010. 5, 6

[21] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3d reconstruction in real-time. In *IV*, 2011. 6

[22] M. Goebl and G. Faerber. A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles. In *IV*, 2007. 2

[23] H. Hirschmueller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30:328–41, 2008. 5

[24] B. K. P. Horn and B. G. Schunck. Determining optical flow: A retrospective. *AI*, 59:81–87, 1993. 6

[25] M. Kaess, K. Ni, and F. Dellaert. Flow separation for fast and robust stereo odometry. In *ICRA*, 2009. 6

[26] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, pages 508–515, 2001. 5, 6

[27] J. Kostkova. Stratified dense matching for stereopsis in complex scenes. In *BMVC*, 2003. 5, 6

[28] R. Kuemmerle, B. Steder, C. Dornhege, M. Ruhnke, G. Grisetti, C. Stachniss, and A. Kleiner. On measuring the accuracy of SLAM algorithms. *Auton. Robots*, 27:387–407, 2009. 2, 5

[29] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*, 2010. 1, 3

[30] S. Morales and R. Klette. Ground truth evaluation of stereo algorithms for real world applications. In *ACCV Workshops*, volume 2 of *LNCS*, pages 152–162, 2010. 1, 3

[31] P. Moreels and P. Perona. Evaluation of features, detectors and descriptors based on 3d objects. *IJCV*, 73:263–284, 2007. 2, 3

[32] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Generic and real-time structure from motion using local bundle adjustment. *IVC*, 27:1178–1193, 2009. 6

[33] Nayar and H. Murase. Columbia Object Image Library: COIL-100. Technical report, Department of Computer Science, Columbia University, 1996. 2, 3

[34] M. Ozuysal, V. Lepetit, and P.Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009. 2, 3

[35] G. Pandey, J. R. McBride, and R. M. Eustice. Ford campus vision and lidar data set. *IJRR*, 2011. 2, 3

[36] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005. 7

[37] T. P. H. B. Rene Ranftl, Stefan Gehrig. Pushing the limits of stereo using variational stereo estimation. In *IV*, 2012. 5, 6

[38] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, 2011. 5, 6

[39] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77:157–173, 2008. 2, 3

[40] A. Saxena, J. Schulte, and A. Y. Ng. Depth estimation using monocular and stereo cues. In *IJCAI*, 2007. 3

[41] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2001. 1, 2, 3, 4, 5

[42] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *IJRR*, 28:595–599, 2009. 2, 3

[43] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart. Towards a benchmark for RGB-D SLAM evaluation. In *RGB-D Workshop*, 2011. 2, 3

[44] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010. 6

[45] M. Werlberger. *Convex Approaches for High Performance Video Processing*. phdthesis, Graz University of Technology, 2012. 5, 6

[46] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *arXiv:1204.1393v1*, 2012. 5, 6

[47] J. yves Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker. *Intel*, 2000. 6

[48] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *DAGM*, pages 214–223, 2007. 6