Mind the Gap – A Benchmark for Dense Depth Prediction Beyond Lidar

Hendrik Schilling[1], Marcel Gutsche[1], Alexander Brock[2], Dane Späth[2], Carsten Rother[2], Karsten Krispin[1]
[1]rabbitAI, Heidelberg [2]Visual Learning Lab, Heidelberg University
firstname@rabbitAI.de firstname.lastname@iwr.uni-heidelberg.de

Abstract

The large interest in autonomous vehicles is a significant driver for computer vision research. Current deep learning approaches are capable of impressive feats, like dense full frame depth prediction from a single image. While impressive results have been achieved, it is not yet clear if they are sufficient for autonomous driving. The problem remains that existing evaluation benchmarks and metrics are not yet capable of fully addressing this issue. This work takes a step towards answering this question. Current evaluation methods are incapable of proving or refuting suitability for potentially hazardous real world situations. This is due to a) the large gaps in the currently used Lidar ground truth data, which cannot test many difficult and relevant cases and b) the global summary metrics used, which are intangible with respect to rigorous performance guarantees. In this work we provide a new benchmark based on commercially available dense light-field depth data, which closes these gaps in the evaluation. We implement domain-specific and interpretable error metrics, which allow for strict assertions over the performance of tested methods. The leaderboard for dense depth prediction is publicly available. The approach is also transferable to other depth estimation tasks. Such stringent evaluations are indispensable when testing and demonstrating performance for potentially hazardous applications like autonomous driving, and are a critical aspect for the assessment of autonomous systems by regulatory bodies as well as for public acceptance.

1. Introduction

In recent years, computer vision has made tremendous progress in solving the challenging computer vision tasks that will eventually make autonomous driving a reality. This progress has been fueled by the ability to train large neural networks as well as the availability of large data sets. In this respect, the significance of publicly available benchmarks should not be underrated. They foster objective and reproducible research, benefiting the research community as well as the industry seeking to implement the results. This work contributes such a benchmark, see Fig. 1 for an example from our evaluation. While there is an abundance of both training data and publicly available benchmarks for semantic computer vision tasks [20] , the situation for depth based computer vision tasks is less promising. To sensibly evalu-

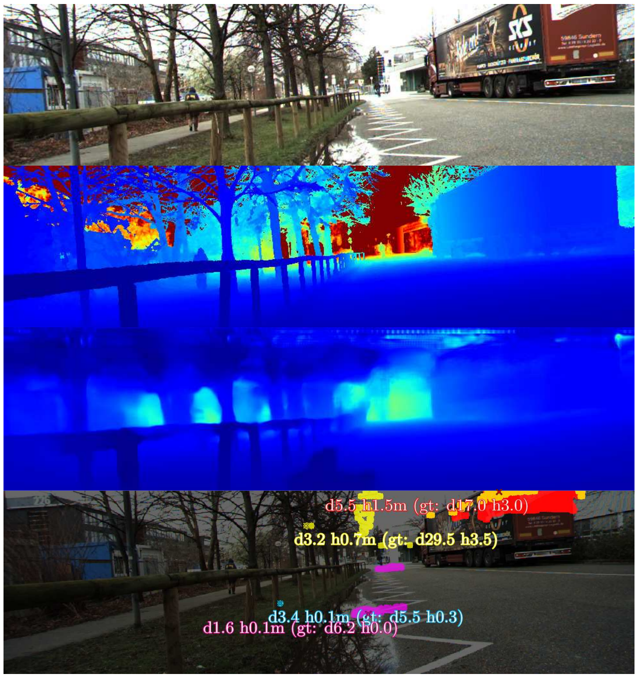
d5.5 h1.5m (gt: d17.0 h3.0)
d3.2 h0.7m (gt: d29.5 h3.5)
d3.4 h0.1m (gt: d5.5 h0.3)
d1.6 h0.1m (gt: d6.2 h0.0)

Figure 1: Example from our new benchmark, from top to bottom: Input image, light-field depth, prediction [24], critical failures (compare Section 9). Closest failure location marked with a cross, with method distance and height (GT in brackets). These represent critical failure cases for autonomous vehicles, which dangerously influence driving behavior due to interference with the drivable corridor, up to 2 m above the street. These critical failures cannot be detected using Lidar, because reflective areas (puddle) and large image heights and/or distances (sky/truck) are missing from Lidar data. Our metrics specifically detect the criticality of these failures, while global metrics like MSE or BadPix fail to determine the influence of errors on autonomous vehicles.

ate deep learning based methods in the autonomous driving scenario, a wide and realistic variety of street level scenes is required. Yet, apart from the well-known KITTI data set [11], all benchmarks are either based on completely synthetic data [8, 32], or limited to a small set of scenes [36, 34], from a domain markedly different from the automotive scenario.

We believe that the main reason for this scarcity of data is the high complexity in acquiring and processing the 3D ground truth required for such data sets. Until recently only Lidar (Light detection and ranging) based acquisition methods were capable of producing this depth information. However, Lidar acquisition has a completely different performance envelope compared to image based depth estimation. The best commercially available Lidar sensors are capable of recording merely 128 points simultaneously (*e.g*. Ouster OS1-128 [28], Velodyne Alpha Prime [39]), and hence have to fall back to scanning a scene by moving or redirecting the Lidar beam. This means that acquiring the 3D information for the dynamic scenarios encountered during autonomous driving is quite challenging as the scene changes during acquisition and there will therefore be mismatches between the image data and the recorded Lidar point clouds. These technical obstacles explain why of the 37 automotive data sets listed by Kang *et al*. [20], 10 provide Lidar data but only two were processed to correctly account for dynamic object motions, specifically KITTI [11] and the HCI data set [21]. But even on the static parts of a scene, the Lidar measurements have many gaps, including very reflective or dark objects, where no valid depth measurements are available. As most depth prediction methods have the potential to work in these situations, the sparse Lidar data is neither sufficient to assess the full potential of image based depth estimation, nor to detect all possible limitations of the predictions. This makes the task of testing and verifying autonomous systems very difficult.

In this work we introduce a benchmark based on passive light-field based dense depth ground truth which has recently become commercially available [30]. The depth data is based on a commercial 17-camera light-field setup. Setup and per-pixel depth data were supplied by rabbitAI[1]. Using this dense depth data for benchmarking we are able to implement a range of improvements over the current Lidar based benchmarking approaches.

In the following we summarize our main contributions:

- A new benchmark with a public leaderboard (rabbitai.de/benchmark), which closes many gaps left by previous approaches.
- Evaluation metrics specific to the domain of autonomous driving which enable stricter assertions with respect to the performance of tested methods.
- A detailed comparison between the previously used Lidar data and the new passive light-field depth.

This benchmark is a step towards strict and interpretable benchmarking for autonomous driving scenarios, and the introduced methodology represents a way of testing and promoting this robustness, for example for regulatory bodies.

2. Related Work

In the following we will introduce previous benchmarks and their data acquisition approaches. Note that we only report benchmarks using real world captures. Image synthesis is in principle able to generate sufficiently realistic imagery, but the modeling of the world to a sufficiently high degree is extremely expensive. Indeed, for realistic content generation many feature films and games rely on 3D scanning methods [27].

In the past, depth ground truth has been acquired with a range of methods, including manual labeling of planes [33] and structured light scanning [34]. Fluorescent UV paint has successfully been utilized for optical flow data sets [5], a method also applicable for depth ground truth. However, all these methods are constrained to static close-range captures and hence cannot provide the range and speed required for dynamic automotive scenarios.

Hence, all current automotive data sets and benchmarks make use of Lidar measurements to acquire depth information. Lidar sensors (short for *Light detection and ranging*) actively scan the scene to determine distances. Compare [20] for an overview of many driving data sets, some of which include depth data.

Two categories of Lidar data sets can be distinguished. Static scene scanning, where a possibly quite slow survey grade Lidar sensor scans a large area, which is then rigidly registered. For this class, dynamic objects need to be handled completely separately, for example by manual fitting of CAD models [26] or using manual annotation of cardboard-style motion [21]. The common problem with these approaches is the extremely labor-intense processing and the limitation to very few classes of dynamic objects, hence some data sets only include the static background and completely ignore dynamic objects for depth estimation, like the Apollo data set [17].

Most automotive data sets that provide depth measurements are based on fast automotive Lidar sensors that have a relatively high scanning rate (10-20Hz) which reduces the skew between camera images and the Lidar measurements. However, the sequential nature of Lidar sensors still introduces significant skew between camera images and Lidar measurements, which must be accounted for. To the best of our knowledge only two data sets are available which perform this post-processing. One is the well-known KITTI data set and benchmark [11], which also incorporates static

[1] rabbitai.de

scene scanning. The second is a recent stereo data set by Yang *et al*. [40], which implements an automatic filtering procedure based on stereo matching – with all the biases that might be introduced. All other data sets simply provide raw Lidar scans with full motion artifacts [29, 6, 25, 37, 2, 31].

None of the data sets address the large gaps which are inevitable due to the measurement principle of Lidar sensors, compare Section 6.

3. Design Goals

In the following we will outline the design goals which governed all decisions for our new benchmark.

Coverage Systematic gaps in the data limit the validity of any conclusions derived from an evaluation, and should therefore be avoided at all costs. Many gaps previously encountered in Lidar based data sets are closed in our new benchmark, see Section 6 for details.

Interpretability Evaluation metrics have limited utility without a way to infer tangible conclusions from them. Global metrics like MSE or SILog [10] do not allow assertions about the suitability of methods for autonomous driving. Metric such as *percentage of obstacles missed at distance x*, as implemented in our benchmark, allow for much stricter performance assessments.

Comparability When comparing methods across several benchmarks it is oftentimes difficult to reach definitive conclusions about the relative performance, as many benchmarks provide totally different imaging characteristics, with their own sets of training data. This makes it difficult to attribute performance differences across benchmarks. Hence, instead of directly using the raw image data, we imitate an established data set, specifically the KITTI imaging pipeline. This has the added benefit of bootstrapping our benchmark with the depth estimation approaches as trained by the respective authors for the original KITTI benchmark.

Updates Over time, a benchmark becomes increasingly outdated. An example for this is the rise of E-Scooters in recent years, which represent a novel hazard that is not present in data obtained before ca. 2018. To enable updates to our benchmark we use the concept of *container submissions*, where submissions are containerized implementations instead of results. We still allow for regular submissions of results, however those are discouraged.

Incentivize Good Submissions An ideal submission matches the following requirements:

- Is a method notably distinct from other submissions.
- Is published in a peer-reviewed conference/journal.
- Has a published implementation.
- Has a containerized algorithm for testing.

The first two aspects are hard requirements, which will only be lifted temporally, *e.g*. to enable a submission to a peer-reviewed venue, and the remaining two aspects will be encouraged via our submission policy.

Data Variety Classically, one would choose a car to capture images from an automotive perspective. However, cars need to follow the rules and flow of traffic, and it is difficult, both for safety and legal reasons, to actively direct a car towards the scenes which are most interesting for an autonomous driving benchmark. To increase the variability in captured scenes and the density of difficult and potentially hazardous scenes for autonomous vehicles we instead opted for mounting the capture setup on a cargo bike, which gives this benchmark a unique perspective for driving situations.

4. Setup

The setup used for this benchmark is a 17 camera light-field setup using Sony IMX253 CMOS sensors and 8mm lenses for a HVOF of around 90°. The resolution of the cameras is 12MP (4096x3000). The setup was mounted on a modified cargo bike, together with additional sensors not relevant for this benchmark, like GNSS receivers. Calibration, recording and depth processing for this benchmark was supplied by rabbitAI [30].

5. Recordings

For the benchmark 9 hours of footage were captured over a period of five weeks in the city of Heidelberg, Germany. From this footage 100 scenes were selected for the actual benchmark, and a further 100 scenes will be released for testing and fine-tuning of submissions.

6. Ground Truth Depth

The ground truth depth data is provided and processed by rabbitAI [30] using the multi-camera setup described in Section 4. The processing includes manual quality control and annotation to provide pixel accurate depth data. In the following we will give a detailed analysis of the different performance characteristics of Lidar depth in comparison with light-field depth used in this benchmark. Note that we compare single shot light-field data to automotive Lidar. Both approaches can be used in a global setting where multiple captures are registered with respect to each other. However, this is even more problematic for dynamic scenes, due to the reasons described earlier.

The measurement characteristics between Lidar and passive light-field depth are fundamentally different, see Table 1

	light-field depth	automotive Lidar
density	**high (0.022°H/V)**	mixed (0.08°H x 0.42°V)
accuracy	depth dependent	**high**
coverage	**full**	limited
range	**unlimited** (see Section 6.2)	40-120m [19]
camera sync	**by design**	skewed
viewpoint	**identical to img**	occlusion artifacts

Table 1: Overview of the characteristics of Lidar and light-field capture for the evaluation of depth prediction.

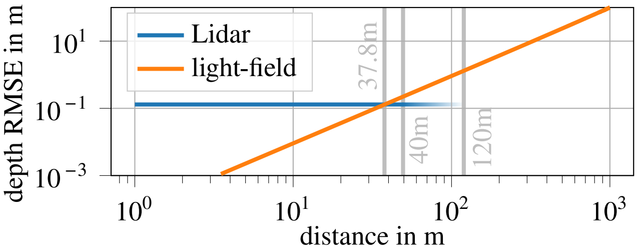

Figure 2: Plot comparing accuracy and completeness of Lidar to the light-field depth, lower is better. While Lidar provides better absolute accuracy from the intersection at 37.8 m, the light-field depth is still usable at much greater distances. For example at 300 meters the light-field depth is accurate to 15 meters, meaning we can be quite certain that objects are at least 285 meters away. At the same time the maximum range of Lidar depends on the reflectivity, and starts to drop some samples at 50 m (street). Behind the max range of 120 m even cars and foliage are dropped [19, 12].

for an overview. The most relevant aspects in the context of this benchmark are range, accuracy and completeness of the captured data. Lidar data has a very constant absolute accuracy, while light-field data is highly depth-dependent. On the other hand Lidar, being an active measurement method, has many issues regarding missing returns which leads to gaps in the measurements. In the following, these three key aspects are analyzed in detail.

6.1. Depth Accuracy versus Depth Range

The accuracy of the Lidar measurements is mostly independent of the distance, although some bias with respect to the surface normal might be present in current data sets [23]. Exact figures on the absolute accuracy are difficult to find, but the standard deviation for different reflectivities has been measured as $0.13\,\mathrm{m}$ [19] for the Velodyne Lidar used by KITTI. Most parameters of the measured objects, like reflectivity lead to a complete loss of data points, but seldom to large errors. On the other hand, the light-field ground truth used in this benchmark is for the most part a passive triangulation based approach which leads to a constant accuracy in disparity space, which induces a strong dependency on the measured distance. For Fig. 2 we assumed a root-mean-square error of 1.5 pixel, which is surpassed by all the state-of-the-art methods on the HCI 4D light-field benchmark [14, 18], including classic approaches not based on deep learning [35]. Note that both Lidar and light-field accuracy should be regarded with a grain of salt, as the Lidar error does merely represent a consistency measure, not an absolute depth error which could be significantly larger [23], while the light-field RMSE is an absolute error measure over a set of benchmarks scenes, but evaluated on synthetic data. However, the exact value for the accuracies does not change the relevant take home message, compare Fig. 2: Lidar does overtake the accuracy of the light-field depth ($37.8\,\mathrm{m}$ for the HDL64E used in KITTI), but also starts to drop data points from as early as $50\,\mathrm{m}$ [23, 12] (street) until it reaches the maximum range ($120\,\mathrm{m}$ for cars and foliage in the Lidar used in KITTI [12]). This means, the range where Lidar is more accurate *and* does not yet drop relevant samples is only between $37.8\,\mathrm{m}$ and $50\,\mathrm{m}$.

We argue that for most automotive applications a fixed accuracy in the disparity space is acceptable because any autonomous driving agent does operate from an egocentric perspective, where any measurement does necessarily entail an uncertainty which increases with the distance from the observer. It is only important that any ground truth data used for evaluations is significantly better[2] than the method in question. As it is highly unlikely monocular depth prediction can achieve the accuracy available to state-of-the-art light-field depth estimation, this requirement can be considered fulfilled for the ground truth data available in our benchmark.

6.2. Depth Range

The range of Lidar measurements depends on the material properties, specifically on the reflectivity of the surface reflecting the emitted light back to the sensor. Manufacturers often state max ranges from $100\,\mathrm{m}$ to $200\,\mathrm{m}$ [28, 39], however actual measurements give lower ranges, *e.g.* $50\,\mathrm{m}$ to $120\,\mathrm{m}$ for reflectivities between 10% and 80% for the Lidar used in KITTI [12]. In comparison, triangulation based passive depth estimation has a fundamentally unlimited range, in the sense that it can still give probable minimum distances for points at infinity, however the accuracy deteriorates as depth increases, compare Fig. 2. In contrast, Lidar drops distant points completely, which means wrong estimates of close objects (*e.g.* hallucinated obstacles) *cannot* be detected from Lidar data, because there are just no valid measurements for these areas. This is highly problematic because without supervision methods actually lean towards hallucinating close objects in those missing areas. This effect is clearly visible for the top scoring methods in the KITTI depth

[2] recommended: one order of magnitude [22]

prediction benchmark. In the context of autonomous driving, close objects are potential obstacles (*e.g.* tree branches, signs), which means a car actually employing such methods might falsely initiate an emergency braking or even start an evasion maneuver which could be hazardous.

6.3. Completeness

A big advantage of the light-field depth used in this benchmark is the completeness of the measurements, while the accuracy varies, depending on the appearance of the object. However, in the context of monocular depth prediction, Lidar often drops samples due to:

- "large" distances, *e.g.* $50\,\mathrm{m}$ at 10% reflectivity, [12]
- strong motion, as most Lidar points are not captured at the same time as the image due to Lidar scanning,
- occlusions due to the change in perspective between Lidar and camera,
- low reflectance [19],
- very specular reflections (car paint, windows, puddles),
- the sparseness of the Lidar measurement and
- the limited vertical field of view.

Figure 5 shows several examples of these effects from our benchmark and from the KITTI data set.

7. A Note on Depth Ambiguity

One open question for both Lidar and light-field depth are areas which are *actually* ambiguous (compared to only appearing ambiguous), like transparent or refractive areas. In these cases the question is which of the multiple possible depth values for a pixel (*e.g.* foreground or refracted object) should define the depth of a pixel value. As this benchmark is focused on autonomous driving, we always choose the closest point. This means the first object which would interact with a virtual camera ray bundle defines the correct depth value as long as it is at all visible.

8. Data Processing Pipeline

As stated in Section 3, to make our benchmark easily comparable we do not use the raw image data for benchmarking but instead imitate the KITTI imaging pipeline.

8.1. Image Processing

Figure 3 is an example from our image processing pipeline, in comparison with a similar scene from KITTI. Starting with the demosaiced rectified center view of the light-field setup, the following steps are performed: Exposure simulation, image distortion, downsampling, blur, remosaicing, demosaicing with simulated KITTI demosaicing filter, undistortion/rectification, crop to final resolution.

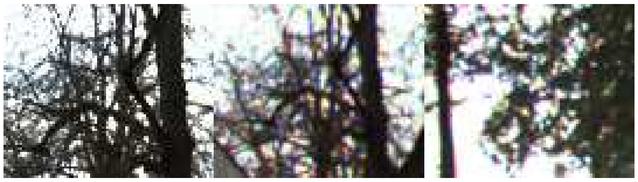

Figure 3: Example processing from our imaging pipeline. From left to right: 1. Clean intermediate image already scaled, exposed, and distorted, 2. output of the pipeline, 3. example patch from KITTI. Note the characteristic color artifacts.

8.2. Depth Data processing

The depth data is initially aligned with the rectified center view of the light-field setup. To align the GT depth with the simulated KITTI images we follow the mappings performed for the image itself, but skip all color based operations, including mosaicing. Also, instead of actually warping the depth, only the image location is warped, resulting in a dense mapping between GT and simulated image. Performing this mapping from the 12MP GT depth to the benchmark depth is then mostly a down-sampling operation. Classical interpolation is problematic on depth maps, because interpolation between distinct objects can lead to depth values which belong to neither foreground nor background. Instead, for every output depth sample we collect all input depth samples which are closer to the desired output point than any other output sample. Then we take the 25% quantile of the depth of these points, to bias towards foreground objects.

9. Evaluation Metrics

As stated in Section 3, the benchmark should be both interpretable and comparable. For this reason we implement well-known global performance metrics used in other benchmarks, including: SILog, sqErrorRel, absErrorRel, iRMSE, scaled by 100 as implemented by KITTI and described by Eigen *et al.* [10]. However, such global metrics only allow for the global rankings of methods, as they cannot be used to deduce the readiness of the tested method for any specific task. For this reason we implement metrics that examine very specific autonomous driving related tasks, which allow for tangible conclusions about the suitability of methods for the tested task. Task specific and geometrically deduced metrics have been used in the past for several depth estimation tasks, from stereo [16] over light-field [15, 18] to optical flow [7].

9.1. Scale Correction

Monocular depth prediction is under-constrained, which often leads to miss-prediction of the absolute scale [10]. We explicitly calculate a scale correction, before performing any evaluation, using a linear model $d_{corr} = \alpha d_{algo} + \beta$ by estimating α, β from pixels on the street mask M via a robust least squares estimate.

9.2. Motivation

All metrics defined below estimate certain failure cases which are relevant to autonomous vehicles. While those metrics are not comprehensive, they check for several very severe failure cases which can lead to dangerous behavior from any vehicle basing decisions on these erroneous depth predictions. We start with the assumption that the two most dangerous scenarios for decisions based on the depth prediction include failure to detect relevant obstacles, which might cause a vehicle to ram the obstacle in question, as well as the hallucination of obstacles, which might cause a vehicle to initiate dangerous collision avoidance maneuvers or perform unwarranted emergency stops. All of these actions are critical hazards which should never occur in regular driving situations.

9.3. Interpretable Metrics

Our error metrics all calculate point sets of erroneous world points Ω_s for every scene $s \in \mathcal{S}$ where $\mathcal{S}$ is the set of all benchmark scenes. We then compare these points sets with a query depth d and calculate the failure ratio E by counting the scenes for which the erroneous point sets contain a point closer to the specified distance:

$$E(\delta) = \frac{1}{|\mathcal{S}|} \cdot \sum_{s \in \mathcal{S}} \begin{cases} 1 & \text{for } |\{r \in \Omega_s \,|\, d(r) < \delta\}| > 0 \\ 0 & \text{else,} \end{cases} \tag{1}$$

which assesses how many failures are encountered at or before a distance threshold δ, where $d(r)$ is the distance of an erroneous point r to the camera plane, compare Fig. 4.

9.4. Street Surface Metric

For the street surface (short *bump*) metric we first reproject all points from D_{GT} and D_A, that lie on the scene specific street mask M, into 3D-space.

We then compute the maximum difference in z-value within sliding windows with size of $1.1\,\text{m} \cdot 1.1\,\text{m}$ along the street plane. The resulting error set for each scene is

$$\begin{aligned} \Omega_s = \{r \in M(D^s_{GT}) \,|\, \varepsilon > 0.07\,\text{m}\}\,, \\ \text{with } \varepsilon = |\Delta_r(D^s_{GT}) - \Delta_r(D^s_A)|\,, \end{aligned} \tag{2}$$

where $M(\cdot)$ is the set of re-projected 3D points on the street mask and $\Delta_r(\cdot)$ computes the range between the 2nd and 98th percentile of street elevation for each sliding window at point r. All windows where ranges $\Delta_r(\cdot)$ between GT and the algorithm deviate by more than $0.07\,\text{m}$ are counted as erroneous.

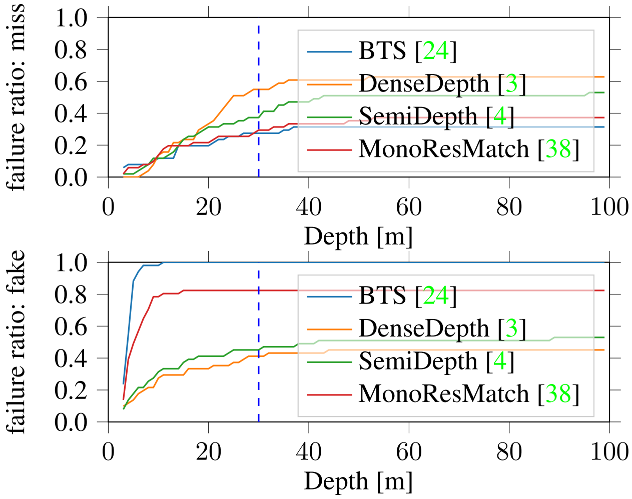

Figure 4: Failure ratios from the *miss* and *fake* metric, lower is better. The *miss* metric (upper) shows failure to detect obstacles close to the visible street surface (like parked cars, bollards). The *fake* metric (lower) detects hallucinated obstacles and displays a clear inversion in the ranking. Safe driving requires *zero* errors at close distances.

9.5. Obstacle Metrics

The obstacle metrics are designed to estimate missing and hallucinated obstacles in the algorithm results. To robustly estimate these, and to limit the metric to relevant obstacles on or close to the street, we always compare two sets of obstacles. A smaller *relevant obstacle set* R determined under stricter thresholds, and a potentially larger *target obstacle set* T, determined with wider thresholds. Failure sets are then determined by removing the wider set from the relevant set.

In the following, we will define the necessary primitives to derive the obstacle metrics. All definitions are based on a depth map D, a height interval H, as well as a street mask $M \subset D$ with an associated expanded street surface S, which is derived using a thin plate spline extrapolation [9] from the projected street surface M^V. We define projections from the depth map: $(\cdot)^V : D \to \mathbb{R}^3$, $(\cdot)^S : D \to \mathbb{R}^2$ and $(\cdot)^I : D \to \mathbb{R}^2$, which project a point from a depth map into the 3D space, 2D position on the street surface and pixel coordinates respectively.

We define the bird-view distance b as

$$b(r) = \min_{d \in M} |r^S - d^S| - \min_{d \notin M} |r^S - d^S|, \tag{3}$$

which evaluates to minus the distance from the street border if r is within M, and to the positive distance from M otherwise. Then $\mathbb{V}$ defines a limited volume above the plane of the street, including off-street areas, as:

$$\mathbb{V} := \left\{ r \in \mathbb{R}^3 | \exists s \in S : (r_z - s_z) \in H \right\}, \tag{4}$$

Metric	Mean30	Miss30	Fake30	MissSt30	FakeSt30	bump30	Avg Scale α	Avg Offset β [m]	SILog[%]	sq_rel[%]	abs_rel[%]	iRMSE [1/km]
SemiDepth[4]	0.26	0.37	0.45	0.2	0.08	0.18	0.96	2.58	38.72	26.53	26.68	38.7
DenseDepth[3]	0.32	0.55	0.41	0.27	0.04	0.35	1.2	3.88	38.37	13.60	24.98	35.48
MonoResMatch[38]	0.47	0.29	0.82	0.25	0.43	0.55	1.03	3.25	43.41	21.98	29.31	48.64
BTS[24]	0.51	0.27	1.00	0.24	0.92	0.14	0.94	2.42	50.96	15.91	27.37	50.28

Table 2: Initial leaderboard, sorted by the mean of our new metrics, evaluated at $30\,\mathrm{m}$ (*mean30*). Note how BTS [24], which is one of the top-performing methods on KITTI comes in last. This is likely due to over-fitting to the KITTI evaluation specifics, while the other methods make use of *e.g.* transfer-learning (DenseDepth [3]) or self/stereo-supervision (SemiDepth [4] and monoResMatch [38]). Global metrics are much higher compared to KITTI due to the completeness of the ground truth, which contains more difficult (far) depth samples. Note that this is only a snapshot, visit rabbitai.de/benchmark for up-to-date results.

where H denotes the relevant height interval above the street. This allows us to define the set of obstacles with a maximum distance R to the street:

$$\mathrm{O} := \{o \in \mathbb{V} \cap D^V | b(o) \le R\}. \tag{5}$$

In addition, we use the closest obstacles operations which selects a *relevant* set of obstacles as those obstacles which are closest to any point in $\mathbb{V}$:

$$C(O) := \{o \in O | \exists r \in \mathbb{V} : o = \arg\min_{c \in \mathbb{O}} |r - c|\} \tag{6}$$

Finally we define an erroneous set as those points from a set R which have no counterpart, within a limited radius of $25\,\mathrm{px}$ in image space, in a target set T:

$$\Omega(R,T) := R \setminus \left\{r \in R | \exists t \in T : |r^I - t^I| \le 25\right\}. \tag{7}$$

Of course this definition only makes sense if R, T are from different disparity sources and without the street pixels ($D_{GT} \setminus M$ and $D_{Algo} \setminus M$). Different parametrization of these sets now yield the final metrics:

	R (source)			T (target)		
metric	src	H	R	src	H	R
miss	$C(GT)$	0.3-2.0	5	A	0.2-2.5	6
fake	$C(A)$	0.3-2.0	5	GT	0.2-2.5	6
missSt	$C(GT)$	0.3-2.0	-0.5	A	0.2-2.5	0.5
fakeSt	$C(A)$	0.3-2.0	-0.5	GT	0.2-2.5	0.5

Specifically *miss* contains obstacles found in the GT that are missing from the algorithm results, *fake* denotes obstacles hallucinated by the algorithm, *missSt* are missing obstacles above the surface of the street mask (*e.g.* boom gate), and *fakeSt* are obstacle above the street surface hallucinated by the algorithm. Note that we use different thresholds for the target set to allow for some absolute movement by the algorithm result. This avoids false errors where some object just outside of the threshold (*e.g.* a car parked on the curb) is just moved by a few centimeters into the threshold by the algorithm.

10. Results

To bootstrap the leaderboard we have taken four monocular depth estimation methods which have publicly available code and pre-trained models, and containerized them: Lee *et al.* [24] (BTS), Alhashim and Wonka [3] (denseDepth), Tosi *et al.* [38] (monoResMatch) and Amiri *et al.* [4] (SemiDepth) a Lidar based extension to Godard *et al.* [13]. The methods were all pre-trained on KITTI by the respective authors, and we report their results using the metrics introduced in Section 9. Table 2 shows the full leaderboard and Fig. 4 shows plots for two of our interpretable metrics. In addition, Fig. 1 and Fig. 5 show a few example results. The full set of results are available on the website (rabbitai.de/benchmark).

The most significant result are the high amount of critical errors, see Fig. 4, which shows failure ratios of over 20% at a distance of $30\,\mathrm{m}$, and still over 5% at $5\,\mathrm{m}$. For safe autonomous navigation, these values need to approach zero. However, there are also positive aspects. Figure 5 and Fig. 1 show visualizations of the algorithm results and the location of critical failures in the image (4th row). We think the failures can mostly be attributed to missing supervision due to reliance on incomplete training data: The shown method was trained with Lidar supervision, and delivers quite convincing results in the lower half of the images, where a lot of supervision was available at training. In the upper part the estimates are very wrong, often hallucinating close objects which would cause emergency braking or collision avoidance maneuvers. The problem is less pronounced in the methods which also use self-supervision (usually stereo) which can provide at least weak supervision in areas where no Lidar GT is available. The failures on the street itself, compare Fig. 1, are also explained by limited supervision, because the Lidar GT used for training cannot provide usable data for highly reflective materials, like the puddle in Fig. 1 or the car paint and shop windows (red rectangles in row *Lidar GT* in Fig. 5). Note that these areas are not amenable to self-supervision, as *e.g.* stereo self-supervision often hinges

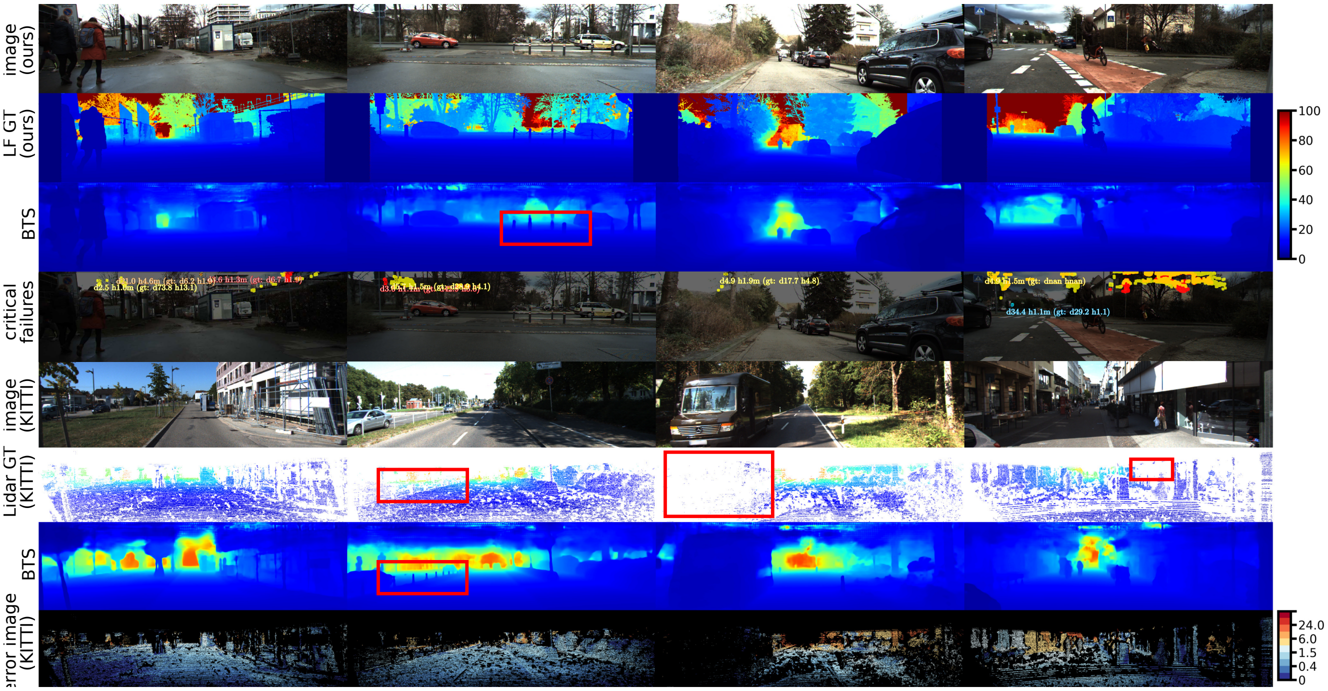

Figure 5: Example scenes from the benchmark and results of BTS [24] (upper half). For comparison also several scenes from KITTI (lower half). The 4th row from the top shows critical failures detected by our metrics : *fakeSt* (red), *fake* (yellow), *miss* (cyan), *missSt* (orange) and *bump* (magenta), and closest failures with wrong distance (d) and height (h) above street, with GT in brackets. Note that all these failures are likely to cause hazardous driving decisions, like triggering unwarranted collision avoidance maneuvers. Although much of the sky is wrong for BTS, which was trained on Lidar data, only the areas highlighted in the 4th row represent dangerous failures which intersect the driving corridor. Note that global error metrics like RMSE are unable to determine which areas are *dangerously wrong* for autonomous vehicles, while our metrics detect specifically those errors which critically affect autonomous driving (*e.g.* by intersecting the driving corridor from street level until 2 m above the ground). Less critical are the fine details missing both in KITTI and BTS, which are available in the light-field GT (red rectangles in the second column). The Lidar GT of KITTI cannot detect many errors, like large parts of the image above the horizon or reflections and small details (red rectangles in the *Lidar GT* row).

on color constancy assumptions.

However, if light-field data is capable of providing reliable test data for these cases it may also be used for training, hence we are eager to see future submissions to our benchmark and their performance improvements on our benchmark. We do not think the solution to the shown problems does necessarily require new network architectures. Many solutions to these challenges are conceivable, from better training data (*e.g.* light-field) over improved training objectives and supervision to explicit handling of the problematic areas, like free space annotation, or manual or automatic completion of existing data sets.

11. Conclusion and Outlook

In summary, this work describes the design of a novel monocular depth prediction benchmark for the scenario of autonomous driving. The benchmark makes use of newly available dense light-field ground truth to implement a much more comprehensive evaluation regime. Specifically, we demonstrate several easily interpretable error metrics, which are capable of detecting critical failures in current depth prediction methods. In addition, we provide a detailed comparison between the classic Lidar based depth ground truth with the novel depth data used in this benchmark. The benchmark is publicly available[3] and will be part of the Robust Vision Challenge, a cross benchmark computer vision challenge aiming to test and promote robust vision methods [1].

This work is a step towards more comprehensive benchmarking, which will improve the robustness of computer vision methods for autonomous driving scenarios. The presented methods may also be useful in demonstrating and promoting this robustness for regulatory bodies and the public. While this work was mostly concerned with depth prediction, the approach can be applied to other vision tasks, like depth completion or stereo image matching.

[3] rabbitai.de/benchmark

References

[1] Robust vision challenge 2020 - eccv2020 workshop. http://www.robustvision.net/. 8

[2] Udacity. public driving dataset. 2017. 3

[3] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 6, 7

[4] Ali Jahani Amiri, Shing Yan Loo, and Hong Zhang. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. *arXiv preprint arXiv:1905.07542*, 2019. 6, 7

[5] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 2

[6] José-Luis Blanco-Claraco, Francisco-Ángel Moreno-Dueñas, and Javier González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research*, 33(2):207–214, 2014. 3

[7] Daniel Alexander Brock and Bernd Jahne. The hci flow metrics: A novel approach for benchmarking optical flow. In *Forum Bildverarbeitung 2018*, page 253. KIT Scientific Publishing, 2018. 5

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2

[9] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 6

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 3, 5

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2

[12] Craig Glennie and Derek Lichti. Static calibration and analysis of the velodyne hdl-64e s2 for high accuracy mobile scanning. *Remote Sensing*, 2, 06 2010. 4, 5

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 7

[14] Katrin Honauer and Ole Johannsen. 4d light field dataset. http://hci-lightfield.iwr.uni-heidelberg.de/, 2016. 4

[15] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*. Springer, 2016. *http://lightfield-analysis.net*. 5

[16] Katrin Honauer, Lena Maier-Hein, and Daniel Kondermann. The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2120–2128, 2015. 5

[17] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. *arXiv: 1803.06184*, 2018. 2

[18] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, Marcel Gutsche, Hae-Gon Jeon, In So Kweon, Alessandro Neri, Jaesik Park, Jinsun Park, Hendrik Schilling, Hao Sheng, Lipeng Si, Michael Strecke, Antonin Sulc, Yu-Wing Tai, Qing Wang, Ting-Chun Wang, Sven Wanner, Zhang Xiong, Jingyi Yu, Shuo Zhang, and Hao Zhu. A taxonomy and evaluation of dense light field depth estimation algorithms. In *Conference on Computer Vision and Pattern Recognition - LF4CV Workshop*, 2017. 4, 5

[19] Maria Jokela, Matti Kutila, and Pasi Pyykönen. Testing and validation of automotive point-cloud sensors in adverse weather conditions. *Applied Sciences*, 9(11):2341, 2019. 4, 5

[20] Yue Kang, Hang Yin, and Christian Berger. Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments. *IEEE Transactions on Intelligent Vehicles*, 4(2):171–185, 2019. 1, 2

[21] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus 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. 2

[22] Daniel Kondermann, Rahul Nair, Stephan Meister, Wolfgang Mischler, Burkhard Güssefeld, Katrin Honauer, Sabine Hofmann, Claus Brenner, and Bernd Jähne. Stereo ground truth with error bars. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 595–610, Cham, 2015. Springer International Publishing. 4

[23] Johann Laconte, Simon-Pierre Deschênes, Mathieu Labussière, and François Pomerleau. Lidar measurement bias estimation via return waveform modelling in a context of 3d mapping. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8100–8106. IEEE, 2019. 4

[24] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 6, 7, 8

[25] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000km: The oxford robotcar dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 3

[26] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2

[27] Jeffrey A Okun and Susan Zwerman. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Taylor & Francis, 2010. 2

[28] Angus Pacala. Introducing the os1-128 lidar sensor. 01 2019. 2, 4

[29] Gaurav Pandey, James R McBride, and Ryan M Eustice. Ford campus vision and lidar data set. *The International Journal of Robotics Research*, 30(13):1543–1552, 2011. 3

[30] rabbitAI. rabbitai. https://rabbitai.de, 2020. 2, 3

[31] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3

[32] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 2

[33] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002. 2

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

[35] Hendrik Schilling, Maximilian Diebold, Bernd Jähne, and Carsten Rother. Trust your model: Light field depth estimation with inline occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[36] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 2

[37] Alex Teichman, Jesse Levinson, and Sebastian Thrun. Towards 3d object recognition via classification of arbitrary object tracks. In *2011 IEEE International Conference on Robotics and Automation*, pages 4034–4041. IEEE, 2011. 3

[38] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019. 6, 7

[39] Velodyne. Velodyne alpha prime vls-128. https://velodynelidar.com/products/alpha-prime/, 03 2020. 2, 4

[40] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 3