OnboardDepth: Depth Prediction for Onboard Systems

Anelia Angelova[1], Devesh Yamparala[2], Justin Vincent[2], Chris Leger[2]

***Abstract*— Depth sensing is important for robotics systems for both navigation and manipulation tasks. We here present a learning-based system which predicts accurate scene depth and can take advantage of many types of sensor supervision. We develop an algorithm which combines both supervised and unsupervised constraints to produce high quality depth and which is robust to the presence of noise, sparse sensing, and missing information. Our system is running onboard in real-time, is easy to deploy, and is applicable to a variety of robot platforms.**

I. Introduction

Predicting scene depth from input imagery is important due to its application to autonomous navigation and manipulation in robotics. Recent work on image-to-depth prediction has demonstrated good quality depth prediction from a monocular camera only, and without additional supervision [1], or by imposing left-right consistency from stereo inputs [2]. However, depth sensors are commonly available, and while they may have missing values, they can be used as supervision. We here propose to take advantage of available supervision, when and if it is available, to obtain more accurate depth prediction, crucial for navigation by autonomous vehicles in the real world.

To that end we develop a visual learning algorithm which combines both supervised and unsupervised sensors to obtain higher quality depth, and which is applicable to a variety of onboard systems regardless of the source of depth sensing (Figure 1). Our approach demonstrates that incorporating unsupervised constraints can additionally improve the supervised learning setup. We observe better, higher quality depth and more reliable results on several datasets. While previous approaches have similarly used sensor supervision [3], our algorithm is specifically designed for practical use: it is real-time, running onboard with a contemporary GPU, is accurate and works with various sources of sensing. Furthermore, we observe that training jointly successfully addresses learning in areas of unknown or missing information. This is also important from a practical standpoint, as sensors readings for supervision may be unreliable, missing or noisy, or may not be accessible in some areas.

We also demonstrate that the proposed method is much more robust to both systematic and random noise. In fact, we find that a moderate amount of random dropout of the sensor is valuable, which makes sense as the sensor is noisy. We further observe more robustness in the presence of systematic noise, i.e., in areas where sensor values are consistently missing (Figure 2). This result may sound surprising in its outcome (Figure 2), but is expected – a well known phenomenon in machine learning is that no meaningful prediction can be made for parts of the domain not covered by any training data [4]. In other words, learning systems are not guaranteed to provide meaningful outputs for these inputs. While prior robotics approaches address this problem with establishing uncertainty on the prediction [5], practical robotics systems will often encounter the problem in which areas of the image or certain types of surfaces, e.g. specular or transparent surfaces, are consistently unknown and return no signal. Thus these areas will output values with zero or very low confidence, which is not particularly useful. We address this problem by combining an approach which is grounded in 3D geometry and uses the underlying geometry during training, and at the same time taking advantage of learning techniques which efficiently extract information from large datasets and available supervision, however sparse.

In summary, this paper presents an onboard system for depth prediction which combines supervised signals with unsupervised constraints which are grounded in the 3D geometry of the world. It provides high quality depth predic-

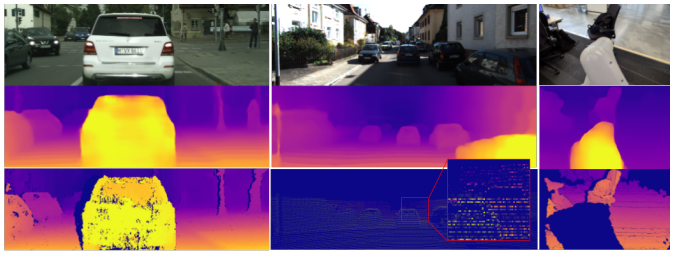

Fig. 1. Learning scene depth for onboard prediction. Top row: input image, middle row - predicted depth, bottom row - sensor depth. Our approach can take advantage of any sensor whenever available: left - an example from the Cityscapes dataset with stereo depth, middle: KITTI dataset with LiDAR ground truth, right: indoor robot manipulation dataset with an IRDS sensor. The prediction covers the full images, whereas the sensor may be very sparse (e.g. KITTI dataset in the middle, with the magnified area for better visibility) or with areas of consistently missing information, e.g. on the surface of the robot arm (bottom right).

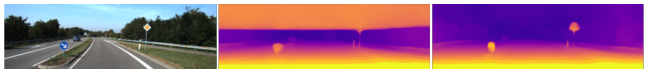

Fig. 2. Depth learning when using standard supervised techniques (middle) may not produce meaningful results for areas of no training data, which is a common problem of applying learning algorithms to real-life systems. Our algorithm counteracts that by combining supervised and unsupervised constraints (right).

[1] Robotics at Google, Research at Google, [2] X, Mountain View, USA. Contact author: anelia@google.com
978-1-7281-3605-9/19/$31.00

tions and runs in real-time and onboard. More specifically, our approach alone takes only 12ms on a Geforce 1080Ti for standard image sizes of 416x128, used in prior depth prediction work. When running onboard our robot, concurrently with other perception modules, it takes 112ms $\pm$ 16ms for image sizes of 640x512 on a GV100 GPU. It is applicable to many sources of sensor supervision, e.g. LiDAR, IRDS and stereo and is much more robust to noise especially for systematically missing values.

II. Related work

Scene depth estimation has been an active research topic due to its importance for navigation and manipulation in robotics [6], [7], [8]. Many previous methods for depth estimation exist, e.g. stereo, active sensing and so on, while learning-based methods have been proposed only recently [8], [9], [10], [11]. In these, a depth estimation function is learned from data, and then depth is predicted from input images. Supervised learning can be supplemented by sensor fusion [12], where a subset of sensor points are additionally available at test time. Such fusion techniques are complementary to ours and can be applied in addition for more robustness.

More interestingly, unsupervised image-to-depth learning has also been proposed recently [1], [13], [2], [14], [15], where the only supervision is obtained from a monocular video. The work of Garg et al. [13] introduced joint learning of depth and ego-motion in a neural based framework. Zhou et al. [1] proposed a neural based approach which is fully differentiable and showed it outperforms prior approaches which used depth sensors as supervision. These works have established the methods for unsupervised depth and ego-motion learning and many subsequent works have improved the initial results in the same monocular setting [14], [16], [17], [18]. Furthermore, learning from stereo inputs has shown success. For example stereo pair videos have been used during training [2], [15], [19], [20] to also produce a single high quality image-based depth estimation. These methods tend to achieve better quality results, due to the extra camera input. The abovementioned learning based approaches, whether supervised, unsupervised or supervised by stereo have demonstrated that learning is a viable approach and an alternative to purely geometric approaches. This is because learning has the opportunity to 'see' a lot of data and thus forms priors from the large amounts of previously observed data, scenes and objects to make a decision. While our approach is related to all of the above, it combines elements of purely geometric constraints with the learning-based setting. That is, it combines both supervised and unsupervised techniques, taking advantage of unsupervised constraints where no supervision is available and vice versa.

Our work belongs to the large class of semi-supervised methods, see [21] for review. Whereas standard techniques assume the input datasets belong to the same domain or (in the case of domain adaptation) at least comply with the same input and output format, our approach is different as it combines inherent scene geometry with learning techniques. In the context of depth prediction, semi-supervised learning has been spearheaded by [3] who use sparse LiDAR depth and apply a combination of supervised and unsupervised losses by enforcing photoconsistency. Their work, however, used an additional stereo input for training. Also related to our work is learning with weak supervision [22], as the approach addresses learning with largely missing supervision information. Furthermore an additional depth sensor can be used during testing for online 'fusion' [12].

III. Main method

In this section we describe the approach. Overview is shown in Figure 3. The input to the system during inference is a single RGB image. The desired output is the depth of the scene corresponding to the image. In order to learn depth, naturally we can apply standard supervised techniques and define a loss

$$L_{gt,I} = min_\theta \sum_i |D^i(\theta) - D^i_a| \tag{1}$$

where the above is the loss per image I and D^i and D^i_a are the predicted and actual depths per pixel, and θ are the learnable parameters. The loss above is naturally summed over all the training images available $L_{gt} = \sum_I L_{gt,I}$.

Importantly, since the ground truth produced by a sensor is noisy and will always have missing values, we need to apply a validity mask and effectively switch off the loss for the missing ground truth values. If we do not do that, the function will try to fit to an incorrect fixed value e.g. 0, which is not desired. Similarly, if during training we wish to eliminate some inputs, e.g. to counteract noisy sensors (which, as seen later, is useful), the same mechanism is used.

$$L_{gt} = min_\theta \sum_I \sum_i M_i |D^i(\theta) - D^i_a| \tag{2}$$

The depth prediction itself is done by a neural network (depth network) which is a dense prediction, fully convolutional network. This choice is made so that it can take advantage of large amounts of training data, including for pre-training.

In our approach we propose to additionally use geometric constrains during training, which are available for free as the supervision is derived from the input and neighboring images. Such geometry-grounded constraints are important to incorporate because the sensor may be noisy or missing in large areas of the image (e.g. Figure 1 middle, right). Crucially, these sensors may introduce systematic errors and consistently fail to produce accurate (or any) results on certain types of surfaces, e.g. specular, reflective, on moving objects etc.

To that end we additionally incorporate unsupervised learning constraints for scene depth, similar to prior works [1], [14], [18], by utilizing the scene geometry and apply photometrics consistency to impose that the scene corresponding to an image, when transformed and projected, should match the next image. The key insight is that unsupervised learning relies on information from images only and thus areas which are often missed by depth sensors will be

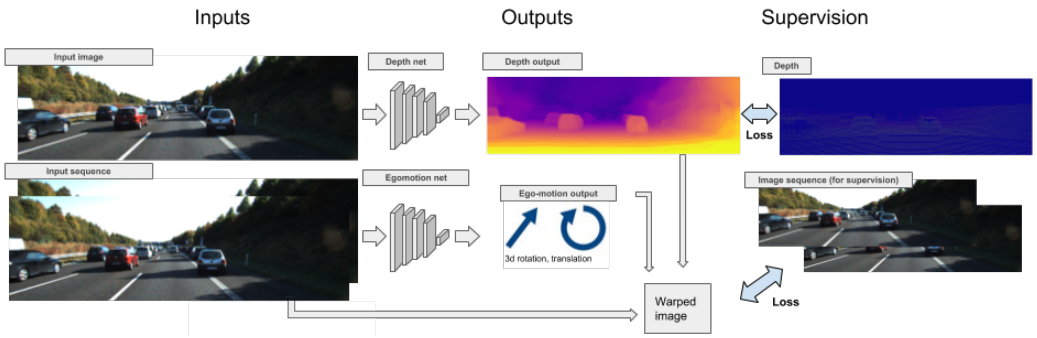

Fig. 3. Main method overview (the depth sensor supervision is hardly visible as it is sparse, top right)

predicted correctly by unsupervised depth methods. Therefore areas which are consistently missed by a sensor will have depth from unsupervised methods which is reconciled throughout learning with supervised losses for areas when supervision is available.

In order to incorporate unsupervised losses, during training two neighboring overlapping images are used and the ego-motion of the vehicle between these two frames is estimated as a sub-product. It is estimated by another deep neural network which is designed to output the rotation and translation of the vehicle in 3D. Thus using the depth of the scene D, and its corresponding point cloud D^s and the 3D ego-motion vector E, one can rotate and translate the point cloud to the next scene $T_E * D^s$, and project it to the image space $I_{warped} = P(T_E * D^s)$ which corresponds to warping the current image to what it should look like at the next frame I_{warped} (here T_E is the transformation derived by ego-motion E, P is the projection matrix, $D^s = P^{-1}K * D$, i.e. inverse projecting or rendering the scene from depth, and K is the calibration matrix, only the focal length is used). This is a differentiable transform as shown by [1] and thus if the next image I_{next} is available, can be added as a loss in the image space. This loss is referred to as photometric consistency, and naturally can be applied to the previous frame as well.

$$L_{unsupervised} = min_\theta \sum_I \sum_i |I^i_{warped}(\theta) - I^i_{next}| \quad (3)$$

where I_{warped} is the current image warped to the next frame. Note that the sensor availability mask is not used here.

Since these principles are both responsible to produce depth they are trained simultaneously (combined via a hyperparamter λ), and a depth output is provided as a result.

$$L_{joint} = L_{gt} + \lambda L_{unsupervised} \quad (4)$$

In addition both the ground-truth (GT) only approach and the joint one have additionally depth smoothness loss and weight regularization loss, as described in [2].

A. Network architectures

Targeting onboard use, we picked a network architecture whose backbone is very efficient. Both depth and ego-motion networks are based on the struct2depth code [18]. More specifically, the depth network uses a ResNet-18 architecture [23]. Due to this choice of main network architecture, we are able to run the algorithm efficiently onboard. Other architectures are also possible, e.g. stemming from the popular U-net or FlowNet architectures. However they do not provide such computational speed as ours.

IV. Experiments

We test the proposed approach on three datasets, both outdoors and indoors and with diverse sensor inputs (LiDAR, stereo, IRDS). We first test on two popular outdoor navigation datasets - KITTI [24] and Cityscapes [25], as well as, on an indoor dataset collected for the purposes of object grasping. The evaluation protocol for depth estimation has been well established and used for both supervised and unsupervised settings [1], [13]. We use the same set of metrics and scripts that many prior authors have used, for example abs_rel error is defined as follows (D^i and D^i_a are the ground truth and predicted depth): abs_rel= $\frac{1}{N}\sum_{i=1}^{N}$ mean($\frac{\|D^i - \hat{D}^i_a\|}{D^i_a}$).

A. Datasets

KITTI. The KITTI dataset [24] is a popular benchmark for developing various algorithms for autonomous driving, such as object detection, tracking, visual odometry, stereo, optical flow and others. It is the most common dataset for depth estimation evaluation, using the sparse LiDAR point cloud as ground truth. KITTI images are resized to 128x416.

Cityscapes. The Cityscapes dataset [25] is a newer urban navigation dataset. It contains higher resolution, more diverse images obtained in populated urban areas of multiple cities. This dataset features many dynamic scenes with moving vehicles and objects. While no ground truth depth sensor is available, we use the stereo depth (disparity) information provided by the dataset. This is in compliance with prior methods [18], [26] which used the sensor and protocol for

Dataset	Ground truth	Train	Test	GT Coverage (percent)
KITTI	LIDAR	39,835	697	4.11%
Cityscapes	Stereo	38,675	1,525	95.78%

TABLE I

DATASETS OVERVIEW: KITTI HAS ONLY 4% GROUND TRUTH COVERAGE.

Method	Range	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Pilzer *et al.*[26]	80m	0.440	6.036	5.443	0.398	0.730	0.887	0.944
Pilzer *et al.*[26]	80m	0.467	7.399	5.741	0.493	0.735	0.890	0.945
Casser *et al.* [18] baseline	80m	0.2049	4.2477	9.0892	0.2543	0.7681	0.9165	0.9626
Casser *et al.* [18] (unsup)	80m	0.2218	5.7374	8.6133	0.2584	0.7738	0.9076	0.9542
Casser *et al.* [18] (unsup+motion)	80m	0.1511	2.4916	7.0237	0.2023	0.8255	0.9372	0.9721
Ours (unsup,)	80m	0.1787	2.9219	8.0368	0.2320	0.7839	0.9259	0.9690
Ours (gt-only, supervised)	80m	0.0952	0.9912	5.9879	0.1583	0.9013	0.9715	0.9891
Ours (joint, supervised)	80m	**0.0936**	**0.9520**	**5.9191**	**0.1563**	**0.9024**	**0.9722**	**0.9897**
Casser *et al.* [18] (unsup)	50m	0.1696	1.7083	6.0151	0.2412	0.7840	0.9279	0.9703
Casser *et al.* [18] (unsup+motion)	50m	0.1529	1.1087	5.5573	0.2272	0.7956	0.9338	0.9752
Ours (unsup)	50m	0.1543	1.5748	4.9975	0.2001	0.8264	0.9435	0.9763
Ours (gt-only, supervised)	50m	0.0787	0.5032	3.4518	0.1283	0.9328	0.9819	0.9934
Ours (joint, supervised)	50m	**0.0773**	**0.4904**	**3.4190**	**0.1267**	**0.9337**	**0.9827**	**0.9937**

TABLE II

EVALUATION OF DEPTH PREDICTION UP TO 50 AND 80 METERS RANGE. CITYSCAPES DATASET. STANDARD DEPTH PREDICTION METRICS ARE SHOWN: FOR COLUMNS IN PURPLE, LOWER IS BETTER, FOR THE BLUE ONES: HIGHER IS BETTER.

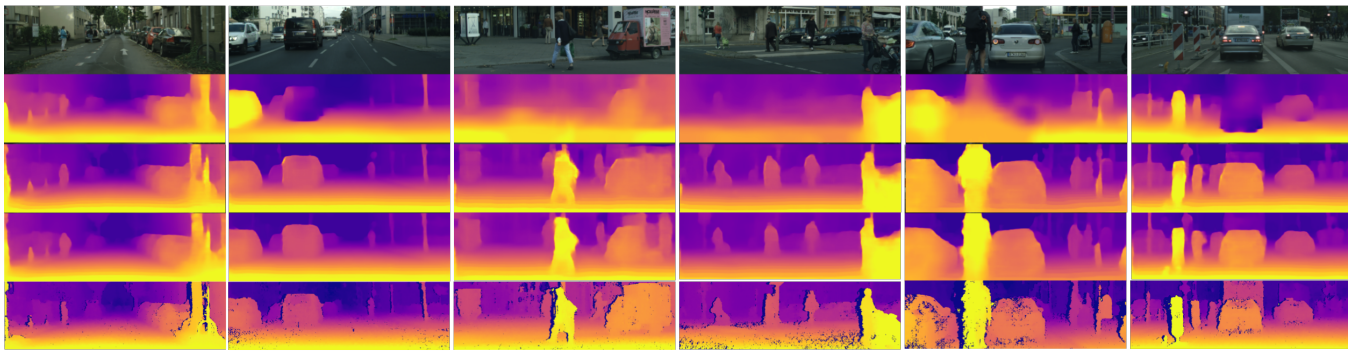

Fig. 4. Depth prediction. From top to bottom, per panel: input image, unsupervised, gt-only, joint, ground truth sensor. Our joint method provides continuous output everywhere (compared to the sensor readings) and is of better quality than the supervised one. Cityscapes dataset.

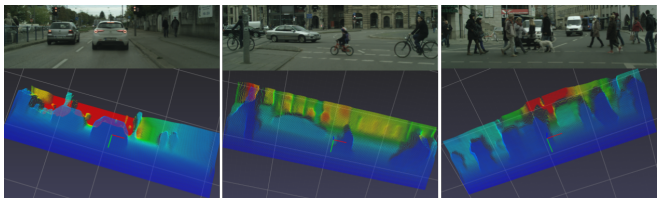

Fig. 5. Point clouds for depth prediction (joint method). Cityscape dataset.

evaluation. This makes our work directly comparable by other approaches which can use the same set of sensors. To unify the experimental setup, the Cityscapes images are cropped centrally and resized to match KITTI image sizes. Table I demonstrates the quantitative differences between the ground truth quality and availability per each dataset. While the Cityscapes dataset provides depth values for about 96% of the pixels in an image, the KITTI dataset is much more sparse providing only about 4 percent of the pixel values.

Indoor robot dataset. We also test the algorithm on an indoor dataset for manipulation, collected by our robot. The depth ground truth is provided by an IR-based depth sensor (IRDS) similar to Microsoft's Kinect [27]. This dataset too has portions of missing values due to areas which are consistently not covered by the sensor, or due to specular, transparent and other challenging surfaces. It has about 11,700 training and 540 test images of size 640x512 (experiments below are conducted on sizes 256x320).

B. Experiments on the Cityscapes dataset: stereo inputs

We first experiment with the Cityscapes dataset in which the supervision comes from stereo. We compare the main algorithms: unsupervised, ground-truth-only (denoted as gt-only) and our proposed joint approach which combines supervised and unsupervised constraints. Table II shows the results and compares to prior methods which have reported results on this dataset. For this dataset no supervised approaches exist, but there are strong performers using motion in addition to depth. In compliance with prior results we evaluate depth up to 80 and 50 meter ranges. As seen, for both ranges, the proposed methods are outperforming the state-of-the-art. While the proposed supervised and joint approaches both use supervision to achieve more accurate depth, such a solution is naturally preferred for a real-world robot system, rather than ignoring this source of more accurate results. We further note that the joint method is consistently better than the ground-truth only one, even though they use the same training data and run within the same computational budget. The improvement is consistent and is also preserved across all metrics, and across all datasets, as also seen later in Table III and Figure 11 for both KITTI and for the indoor robot dataset.

Figure 4 provides qualitative results on Cityscapes. Both supervised methods are of very good quality and more accurate than the unsupervised one. The learned approaches, compared to the raw supervision, also have the advantage of providing values anywhere, whereas the ground truth may have missing values. The joint method is better, also qualitatively, due to occasional erroneous high intensity values for the gt-only method. This is observed at object boundaries, which may be due to sensor errors in these areas. Figure 5 shows the predicted depth in a point cloud representation, where we can see that objects (including moving ones) are correctly placed in the scene. We further note that our approach, being semi-supervised, is able to adequately predict depth of moving objects (Figure 4), which is a well known deficiency for unsupervised methods [28]. In conclusion, we observe that the jointly-supervised approach outperforms the supervised-only one quantitatively and qualitatively. Both achieve better accuracy than other methods and are better suited for onboard depth estimation. With real-time runtime they are particularly suitable in practice.

C. Experiments on KITTI: LiDAR inputs

We further test the approach on the KITTI dataset, where the ground truth is provided by a LiDAR sensor and is sparse. Table III shows the results on the KITTI dataset when evaluating on range up to 80 meters. For this dataset there are both supervised and unsupervised prior approaches to compare to. Our results are compared to prior works, most of which are unsupervised learning methods; they are technologically more advanced than (older) supervised methods and achieve better performance. We here too observe improved accuracy by both the gt-only and the joint approach, and similarly to the Cityscapes dataset, the joint method outperforms the gt-only one. At the bottom of the table we also compare to results which are in a similar setting to ours, but which use additional stereo inputs during training [3]. As seen, our method obtains comparable results. Furthermore, we include comparison with a method that obtains and uses an additional depth sensor at test time [12]. While this is not directly comparable to ours, it is included for completeness, and as seen outperforms all methods. We also note that none of the prior methods we are aware of, are designed for faster speeds or report their runtimes, e.g. [3] used much more powerful networks e.g. ResNet50, which are known to be very slow.

We can conclude here that our method is competitive to prior work on this benchmark, and at the same time has the advantage of real-time execution.

The LiDAR sensor for this dataset has large missing areas for the top half of the image (Figure 6). When comparing our joint method to learning with gt-only we observe that the results are quite close quantitatively. However, due to large blind spots for the sensor, the gt-only technique, despite obtaining accurate results in the supervised region does not generalize very well to other regions. This is visualized in Figure 6 which shows that areas at the top of the image, where no LiDAR points are available in any image, are predicted inaccurately for the gt-only method, whereas our joint method provides much better depth prediction in the whole image. At the same time the quantitative results are close because evaluation is done with respect to the LiDAR sensor on test images, which are missing any values in the top half of the image. That is, the sensor has a blind spot and thus cannot evaluate the methods well in these areas. Our results in the next section (Section IV-D), show similar results when systematically ignoring sensor readings. In these cases too, our proposed joint method is much more accurate and adequate than the gt-only approach.

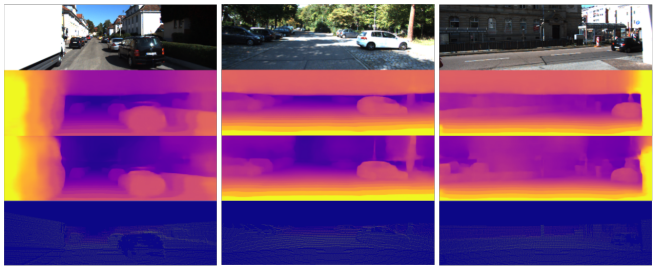

Fig. 6. Quality comparisons between ground-truth only (second row) and our joint approach (third row). KITTI dataset. Our proposed method provides much more accurate results for the top areas of the image, which are not covered by the sensor. The LiDAR sensor is in the bottom row and is very sparse (see also Figure 1 for a magnified example sensor image).

As mentioned earlier, these largely inaccurate values for the ground truth-only method are due to consistently missing data during training. This is also easy to see, given the optimization criteria imposed. Namely, the model is required to fit the ground truth well (its loss and in our case its evaluation is only estimated for these areas of the input), therefore there really is no mechanism to require the system to output specific or consistent values in the areas where information is missing. When measuring the accuracy of the approach, trained on KITTI, but evaluated on an out-of-sample dataset, e.g. the Cityscape dataset, where evaluations are done with respect to another sensor (stereo), we can confirm the above findings too. Figure 7 shows the performance of the model trained on the KITTI dataset, when evaluated on KITTI and Cityscapes. While we should not have high expectations in accuracy for another dataset, this clearly demonstrates the robustness of our approach and the overfitting that the ground-truth-only approach exhibits. At the same time we observe the consistent improvements of the joint method over the gt-only on KITTI itself, as training progresses. As seen, the ground-truth-only method exhibits notable overfitting tendencies and despite seamingly accurate results when evaluated on KITTI does not perform well on an out-of-sample dataset.

Here, while in general uncertainty values for predictions are helpful, our joint approach is much more usable in practice because it provides reliable depth estimates for large missing areas, instead of providing areas with unknown or highly uncertain value. The joint approach provides values anywhere especially in areas where no sensor data is available. This makes it more relevant in practice.

Method	Supervised?	Additional Use?	Cap	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	-	-	80m	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Eigen *et al.*[8] Coarse	GT Depth	-	80m	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al.*[8] Fine	GT Depth	-	80m	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.*[9]	GT Depth	-	80m	0.201	1.584	6.471	0.273	0.68	0.898	0.967
Zhou *et al.*[1]	-	-	80m	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang *et al.*[16]	-	-	80m	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Yang *et al.*[29] (Lego)	-	Motion	80m	0.162	1.352	6.276	0.252	0.783	0.921	0.969
Yin *et al.*[17] (GeoNet)	-	Motion	80m	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Wang *et al.*[30] (DDVO)	-	-	80m	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Casser *et al.*[18]	-	Motion	80m	0.1412	1.0258	5.2905	0.2153	0.8160	0.9452	0.9791
Godard *et al.*[23]	-	-	80m	0.133	1.158	5.370	0.208	0.841	0.949	0.978
Yang *et al.*[19]	-	-	80m	0.137	1.326	6.232	0.224	0.806	0.927	0.973
Yang *et al.*[19]	-	Motion	80m	0.131	1.254	6.117	0.220	0.826	0.931	0.973
Ours (gt-only, sup.)	GT Depth	-	80m	0.1173	0.7810	4.7069	0.1912	0.8615	0.9565	0.9823
Ours (joint, sup.)	GT Depth	-	80m	**0.1159**	**0.7667**	**4.6652**	**0.1892**	**0.8618**	**0.9575**	**0.9834**
Kuznietzov *et al.*[3] unsup.	-	Stereo	80m	0.308	9.367	8.700	0.367	0.752	0.904	0.952
Kuznietzov *et al.*[3] sup.	GT Depth	Stereo	80m	0.122	0.763	4.815	0.194	0.845	0.957	0.987
Kuznietzov *et al.*[3] semi sup.	GT Depth	Stereo	80m	0.113	**0.741**	4.621	**0.189**	0.862	0.960	0.986
Ma and Karaman [12] semi sup.	GT Depth	-	80m	0.208	-	6.266	-	0.591	0.900	0.962
Ma and Karaman [12] semi sup.	GT Depth	Online Depth	80m	**0.073**	-	**3.378**	-	**0.935**	**0.976**	**0.989**

TABLE III

EVALUATION OF DEPTH PREDICTION. KITTI DATASET. COMPARISON TO PREVIOUS MONOCULAR METHODS. A SEMI-SUPERVISED APPROACH WHICH USES STEREO [3] AND ONE WHICH USES DEPTH INFORMATION AT TEST TIME [12] ARE ALSO COMPARED AT THE BOTTOM. PURPLE COLUMNS - LOWER IS BETTER, BLUE COLUMNS - HIGHER IS BETTER.

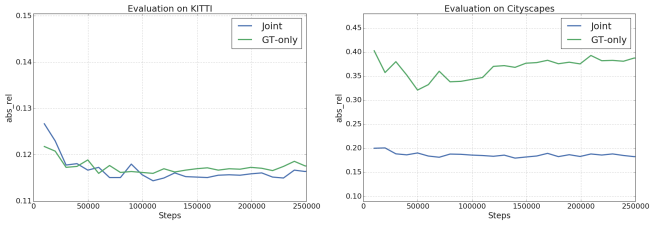

Fig. 7. Comparison between depth prediction results for the gt-only and joint methods on the test set, while training on KITTI. Evaluating on KITTI (left) and on the Cityscapes dataset (right). While the differences are small the joint method consistently improves performance on KITTI. When evaluating on Cityscapes (right), it is seen that the joint method is much more robust, which is primarily due to incorrectly predicted values in the top areas of the image.

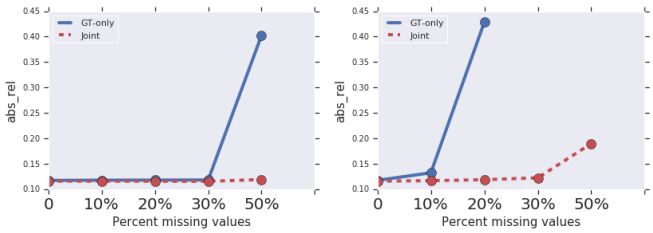

Fig. 8. Performance with introduction of noise in the form of missing values. Random noise (left), consistent noise (right). As seen, our joint approach is more robust in both cases, and especially so for the consistent noise case. The absolute error metric is shown. KITTI dataset.

D. Experiments with missing sensor values

In this section we experiment with the KITTI dataset by introducing experimentally noise in the sensor in the form of missing values. We here test if the joint model and the ground-truth only one are robust when a subset of the images, or large areas in the images miss supervision. The following two sources of missing values are introduced:

Consistently missing values. Consistently missing values are introduced by removing the sensor signal from contiguous areas in the image before training. For the purpose of the experiment, we remove a rectangular area to the right of the image, which covers the image top to bottom; in particular 10, 20, 30, and 50 percent of the image are removed. While simple, this experiment intends to test a scenario in which the sensor is missing frequently or always in some areas of the image, as is the case with KITTI (Section IV-C) and with the indoor dataset (Section IV-E).

Values missing at random. Missing values can also be introduced by removing them at random locations in the image. We do that at several levels, effectively reducing 10, 20, 30 and 50 percent of ground truth values. Since this data has about 4 percent coverage for ground truth, this reaches about 2 percent ground truth values.

We note that each sensor itself has additional sources of noise, e.g. as evident in some of our visualizations (e.g. Figure 4, bottom), additional noise can be further introduced to the sensor measurements.

Figure 8 summarizes the results of training with removing portions of the available ground truth data by either removing values consistently, or at random, as described above. While the proposed joint model is quite robust to both types of noise, the gt-only model quickly deteriorates for systematic noise and is not as robust as the joint model for random noise. Figure 8 (left) shows the results of training a model in the case of randomly removed values, i.e., 10%, 20%, 30%, 50% pixels are removed at random from the ground truth values. As seen, with the exception of very large noise, 50%, where the ground truth model is not performing well, both models are robust to this noise, with the joint model performing consistently better for all values. Furthermore, we find that our joint model performs a little better when trained with some random noise. That is, moderate amounts of random noise makes the model more robust. We observed the same behavior when testing on out of sample data e.g. Cityscapes, where models trained with random noise perform

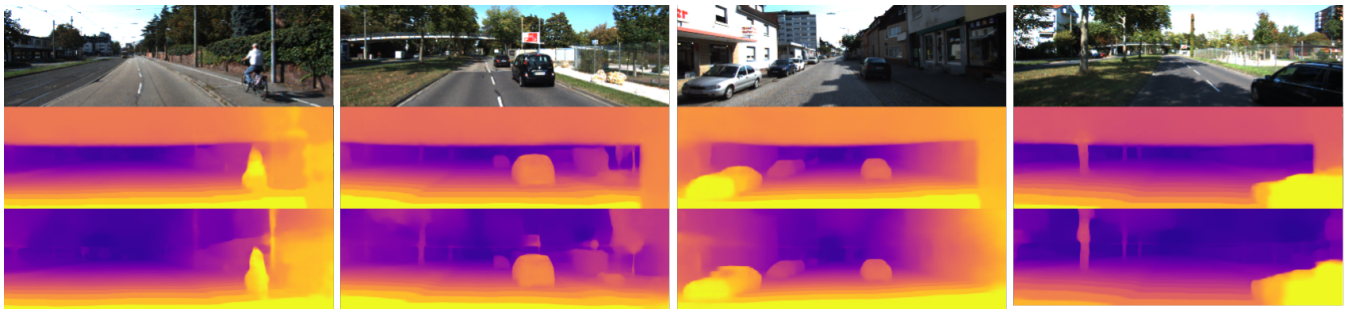

Fig. 9. Quality comparisons between ground-truth only (middle row) and our joint approach (bottom row), when consistent noise is introduced on the right side of the image. KITTI dataset. Our proposed method provides much more accurate results for areas not covered by the sensor.

Fig. 10. Examples of depth prediction from the indoor robot dataset.

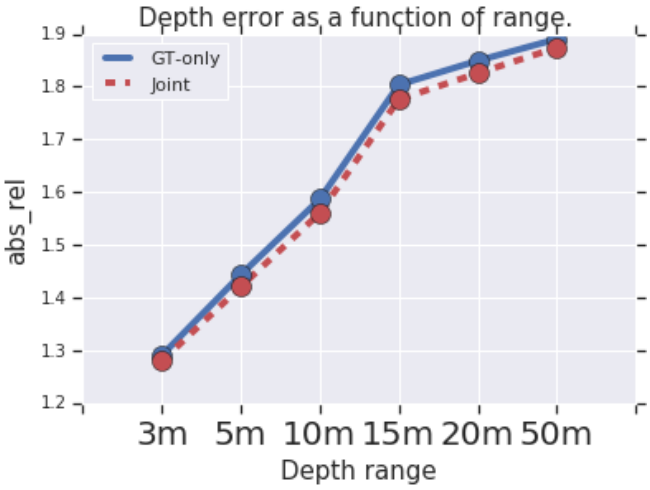

Fig. 11. Quality comparisons between ground-truth only and our joint approach for the Robot arm dataset. Results are shown as a function of range, where the error is evaluated for both methods up to that specific range. The joint approach is consistently outperforming its supervised counterpart.

better, than the model trained without noise. When the noise is on the large size, e.g. 30 to 50 percent then some small deterioration is observed.

Figure 8 (right) shows the results of training a model in the case of consistently missing values, i.e. when a portion of pixels on the side of the image are removed. As seen, systematic errors are harder to overcome than randomly missing values. We here see that the joint model is again much more robust, degrading performance gracefully, whereas the gt-only model is not able to converge to good results for any values of 20% or above. Figure 9 visualizes example depth estimation of both models with consistently missing values, in which the performance degradation is obvious for the gt-only model (at the right side of the image). This shows similar behavior to issues predicting depth in missing sensor areas when training on the full KITTI dataset as seen in Section IV-C. While both supervised and jointly-supervised approaches are immune to portions of randomly missing values, the joint approach is much more robust to systematically missing values, which is a common scenario for physical depth sensors. Skipping values at random (which can easily be accomplished in practice) provides a sense of regularization by obtaining a more robust solution.

E. Experiments for a robot arm: IRDS sensor

We apply the proposed algorithm for the purposes of an arm end-effector manipulating objects. Here the goal is to obtain depth for table top objects for grasping. This experiment is done in real environments, on naturally occurring office spaces, which have not been specially set up or modified for the experiment. The ground truth depth is provided by an IRDS sensor which may have some missing values. The sensor produces estimates within 20 meters and the most accurate estimates within 10 meters. Its behavior is in-between the sensors from the previous two experiments, as the sensor readings are dense, but areas of the image may be missing, for example portions to the left of the image are often not available; the robot arm itself is reflective, which often is lacking sensor values (Figure 1).

Figure 10 shows the estimated depth for the robot arm. While the robot is intended to work on table-top surfaces, we collected images from afar as well in order to have a more challenging and diverse test set. Similar to other datasets, here we have continuous depth estimation, providing depth everywhere, closing the gaps of systematic missing values. Figure 11 shows depth prediction error as a function of range. Depth is evaluated up to 3, 5, 10, 20 and up to 50 meters where available by the sensor, whereas training is done on ground truth up to 15 meters. Here too, we can observe that the joint model outperforms the gt-only one consistently. Figure 12 further shows comparison of the joint and ground-truth only algorithms. As seen, they perform similarly. The

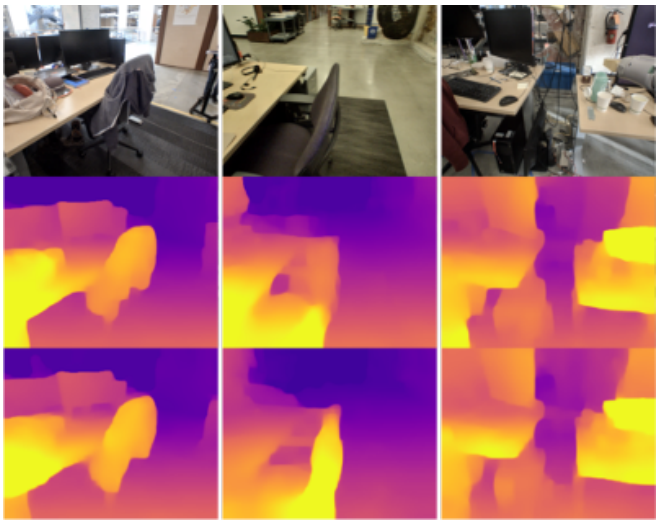

Fig. 12. Examples of depth estimation from the indoor robot dataset. Input image (top), depth prediction from supervised-only learning (middle), depth prediction from joint learning (bottom).

gt-only algorithm is sometimes more prone to errors to the left of the image where there are consistently missing values (right images). While the algorithm is intended to work at short ranges (up to 15 meters) we can also see that it is able to get depth beyond desks where objects are found. Here too, for the indoor robot arm dataset with an IRDS sensor, we see successful depth prediction and also consistently better performance of the joint model compared to the gt-only on all ranges.

V. Conclusions and Future Work

This paper presents a joint approach combining supervised learning techniques with unsupervised constraints, which obtain better onboard depth prediction. These constraints are geometrically grounded and are available for free and are beneficial for practical applications. Indeed we see consistent improvements over the supervised-only solution on three datasets, with a variety of supervisory sensors and sparsity levels, when compared over the same dataset and the same inference budget. It also produces a much more accurate solution for areas of missing values and is a more robust approach. Our approach is efficient and is deployed onboard a robot platform. As future work we will consider training across datasets, where one can consider using a global scale per object as a prior e.g. a human or a soda can will have respective sizes which are globally consistent in the world, regardless of the dataset. Furthermore we would like to combine this inference with semantics, namely using the object and class information, as well.

Acknowledgements

We thank Sean Kirmani for his help with running onboard and members of both X and Robotics at Google teams.

References

[1] T. Zhou, M. Brown, N. Snavely, and D. Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," *CVPR*, 2017.
[2] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," *CVPR*, 2017.
[3] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *CVPR*. IEEE, 2017.
[4] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.
[5] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *CVPR*, 2018.
[6] A. Saxena, J. Schulte, and A. Ng, "Depth estimation using monocular and stereo cues," in *International Joint Conference on Artifical Intelligence*, 2007.
[7] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *CVPR*, 2010.
[8] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NIPS*, 2014.
[9] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *PAMI*, 2015.
[10] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," *3DV*, 2016.
[11] V. Guizilini and F. Ramos, "Unpaired learning of dense visual depth estimators for urban environments," in *Proceedings of The 2nd Conference on Robot Learning, PMLR*, 2018.
[12] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *ICRA*, 2018.
[13] R. Garg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *ECCV*, 2016.
[14] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints," in *CVPR*.
[15] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "DeMoN: Depth and Motion Network for Learning Monocular Stereo," *CVPR*, 2017.
[16] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," *arXiv:1711.03665*, 2017.
[17] S. J. Yin, Z., "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," *CVPR*, 2018.
[18] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," *AAAI*, 2019.
[19] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Every Pixel Counts: Unsupervised Geometry Learning with Holistic 3D Motion Understanding," *CoRR:1806.10556*, 2018.
[20] H. Zhan, R. Garg, C. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," *CVPR*, 2018.
[21] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2006.
[22] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, 2017.
[23] C. Godard, O. M. Aodha, and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," *CoRR:1806.01260*, 2018.
[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.
[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
[26] A. Pilzer, D. Xu, M. M. Puscas, E. Ricci, and N. Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," *3DV*, 2018.
[27] A. Wilson and H. Benko, "Combining multiple depth cameras and projectors for interactions on, above, and between surfaces," *ACM symposium on User interface software and technology*, 2010.
[28] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Unsupervised monocular depth and ego-motion learning with structure and semantics," in *CVPR VOCVALC Workshop*, 2019.
[29] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," *CVPR*, 2018.
[30] C. Wang, J. Buenaposada, R. Zhu, and S. Lucey, "Learning Depth from Monocular Videos using Direct Methods," *CVPR*, 2018.