On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks

HyunJun Jung∗1, Patrick Ruhkamp∗1,2, Guangyao Zhai1, Nikolas Brasch1, Yitong Li1, Yannick Verdie1,3, Jifei Song3, Yiren Zhou3, Anil Armagan3, Ales Leonardis3, Nassir Navab1, Benjamin Busam1,2

1 Technical University of Munich, 2 3Dwe.ai, 3 Huawei Noah’s Ark Lab, 4 Siemens AG, ∗ Equal Contribution
hyunjun.jung@tum.de, p.ruhkamp@tum.de, guangyao.zhai@tum.de, b.busam@tum.de

Abstract

Learning-based methods to solve dense 3D vision problems typically train on 3D sensor data. The respectively used principle of measuring distances provides advantages and drawbacks. These are typically not compared nor discussed in the literature due to a lack of multi-modal datasets. Texture-less regions are problematic for structure from motion and stereo, reflective material poses issues for active sensing, and distances for translucent objects are intricate to measure with existing hardware. Training on inaccurate or corrupt data induces model bias and hampers generalisation capabilities. These effects remain unnoticed if the sensor measurement is considered as ground truth during the evaluation. This paper investigates the effect of sensor errors for the dense 3D vision tasks of depth estimation and reconstruction. We rigorously show the significant impact of sensor characteristics on the learned predictions and notice generalisation issues arising from various technologies in everyday household environments. For evaluation, we introduce a carefully designed dataset1 comprising measurements from commodity sensors, namely D-ToF, I-ToF, passive/active stereo, and monocular RGB+P. Our study quantifies the considerable sensor noise impact and paves the way to improved dense vision estimates and targeted data fusion.

1. Introduction

Our world is 3D. Distance measurements are essential for machines to understand and interact with our environment spatially. Autonomous vehicles [23, 30, 50, 58] need this information to drive safely, robot vision requires distance information to manipulate objects [15, 62, 72, 73], and AR realism benefits from spatial understanding [6, 31].

A variety of sensor modalities and depth predic-

Figure 1. Other datasets for dense 3D vision tasks reconstruct the scene as a whole in one pass [8, 12, 56], resulting in low quality and accuracy (cf. red boxes). On the contrary, our dataset scans the background and every object in the scene separately a priori and annotates them as dense and high-quality 3D meshes. Together with precise camera extrinsics from robotic forward-kinematics, this enables a fully dense rendered depth as accurate pixel-wise ground truth with multimodal sensor data, such as RGB with polarization, D-ToF, I-ToF and Active Stereo. Hence, it allows quantifying different downstream 3D vision tasks such as monocular depth estimation, novel view synthesis, or 6D object pose estimation.

tion pipelines exist. The computer vision community thereby benefits from a wide diversity of publicly available datasets [23, 51, 52, 57, 60, 61, 65], which allow for evaluation of depth estimation pipelines. Depending on the setup, different sensors are chosen to provide ground truth (GT) depth maps, all of which have their respective advantages and drawbacks determined by their individual principle of distance reasoning. Pipelines are usually trained on the data without questioning the nature of the depth sensor used for supervision and do not reflect areas of high or low confidence of the GT.

Popular passive sensor setups include multi-view stereo cameras where the known or calibrated spatial relationship

1 dataset available at https://github.com/Junggy/HAMMER-dataset
between them is used for depth reasoning [51]. Corresponding image parts or patches are photometrically or structurally associated, and geometry allows to triangulate points within an overlapping field of view. Such photometric cues are not reliable in low-textured areas and with little ambient light where active sensing can be beneficial [52, 57]. Active stereo can be used to artificially create texture cues in low-textured areas and photon-pulses with a given sampling rate are used in Time-of-Flight (ToF) setups either directly (D-ToF) or indirectly (I-ToF) [26]. With the speed of light, one can measure the distance of objects from the return time of the light pulse, but unwanted multi-reflection artifacts also arise. Reflective and translucent materials are measured at incorrect far distances, and multiple light bounces distort measurements in corners and edges. While ToF signals can still be aggregated for dense depth maps, a similar setup is used with LiDAR sensors which sparsely measure the distance using coordinated rays that bounce from objects in the surrounding. The latter provides ground truth, for instance, for the popular outdoor driving benchmark KITTI [23]. While LiDAR sensing can be costly, radar [21] provides an even sparser but more affordable alternative. Multiple modalities can also be fused to enhance distance estimates. A common issue, however, is the inherent problem of warping onto a common reference frame which requires the information about depth itself [27, 37]. While multi-modal setups have been used to enhance further monocular depth estimation using self-supervision from stereo and temporal cues [25, 60], its performance analysis is mainly limited to average errors and restricted by the individual sensor used. An unconstrained analysis of depth in terms of RMSE compared against a GT sensor only shows part of the picture as different sensing modalities may suffer from drawbacks.

Where are the drawbacks of current depth-sensing modalities - and how does this impact pipelines trained with this (potentially partly erroneous) data? Can self- or semi-supervision overcome some of the limitations posed currently? To objectively investigate these questions, we provide multi-modal sensor data as well as highly accurate annotated depth so that one can analyse the deterioration of popular monocular depth estimation and 3D reconstruction methods (see Fig. 1) on areas of different photometric complexity and with varying structural and material properties while changing the sensor modality used for training. To quantify the impact of sensor characteristics, we build a unique camera rig comprising a set of the most popular indoor depth sensors and acquire synchronised captures with highly accurate ground truth data using 3D scanners and aligned renderings. To this end, our main contributions can be summarized as follows:

1. We question the measurement quality from commodity depth sensor modalities and analyse their impact as supervision signals for the dense 3D vision tasks of depth estimation and reconstruction.

2. We investigate performance on texture-varying material as well as photometrically challenging reflective, translucent and transparent areas where learning methods systematically reproduce sensor errors.

3. To objectively assess and quantify different data sources, we contribute an indoor dataset comprising an unprecedented combination of multi-modal sensors, namely I-ToF, D-ToF, monochrome RGB+P, monochrome stereo, and active light stereo together with highly accurate ground truth.

2. Related Work

2.1. Geometry from X

A variety of sensor modalities have been used to obtain depth maps. Typical datasets comprise one ground truth sensor used for all acquisitions, which is assumed to give accurate enough data to validate the models:

**Stereo Vision.** In the stereo literature, early approaches [51] use a pair of passive cameras and restrict scenes to piece-wise planar objects for triangulation. Complex setups with an industrial robot and structured light can yield ground truth depth for stereo images [1]. Robots have also been used to annotate keypoints on transparent household objects [36]. As these methods are incapable of retrieving reliable depth in textureless areas where stereo matching fails, active sensors are used to project patterns onto the scenes to artificially create structures. The availability of active stereo sensors makes it also possible to acquire real indoor environments [52] where depth data at missing pixels is inpainted. Structure from motion (SfM) is used to generate the depth maps of Sun3D [65] where a moving camera acquires the scenes and data is fused ex post. A temporally tracked handheld active sensor is further used for depth mapping for SLAM evaluation in the pioneering dataset of Sturm et al. [57]. While advancing the field, its depth maps are limited to the active IR-pattern used by its RGB-D sensor.

**Time-of-Flight Sensors.** Further advances in active depth sensing emphasize ToF more. Initial investigations focus on simulated data [26] and controlled environments with little ambient noise [54]. The broader availability of ToF sensors in commercial products (e.g. Microsoft Kinect series) and modern smartphones (e.g. I-ToF of Huawei P30 Pro, D-ToF in Apple iPhone 12) creates a line of research around curing the most common sensor errors. These are multi-path interference (MPI), motion artefacts and a high level of sparsity and shot noise [27]. Aside of classical active and passive stereo, we therefore also include D-ToF and I-ToF modalities in all our experiments.
To extract highly accurate geometry, we design a multi-stage acquisition process. At first, 3D models are extracted with structured light 3D scanners (a). Scene objects (b) and mounted sensor rig (b) are calibrated towards a robot for accurate camera pose retrieval [61]. A motion trajectory is recorded in gravity compensation mode (d) and repeated to record synchronized images of all involved sensors (e). A partial digital twin of the 3D scene (f) is aligned to small (g) and larger (h) objects to retrieve an entire in silico replica of the scene which can be rendered from the camera views of each sensor used (i) which results in highly accurate dense depth maps that enable investigations of individual sensor components.

**Polarimetric Cues.** Other properties of light are used to indirectly retrieve scene surface properties in the form of normals for which the amount of linearly polarized light and its polarization direction provide information, especially for highly reflective and transparent objects [17, 29]. Initial investigations for shape from polarization mainly analyse controlled setups [3, 18, 53, 70]. More recent approaches investigate also sensor fusion methods [28] even in challenging scenes with strong ambient light [60]. We consequently also acquire RGB+P data for all scenes.

**Synthetic Renderings.** In order to produce pixel-perfect ground truth, some scholars render synthetic scenes [40]. While this produces the best possible depth maps, the scenes are artificially created and lack realism, causing pipelines trained on Sintel [7] or SceneFlow [40] to suffer from a synthetic-to-real domain gap. In contrast, we follow a hybrid approach and leverage pixel-perfect synthetic data from modern 3D engines to adjust highly accurate 3D models to real captures.

### 2.2. Monocular Depth Estimation

Depth estimation from a single image is inherently ill-posed. Deep learning has enabled this task for real scenes. **Supervised Training.** Networks can learn to predict depth with supervised training. Eigen et al. [14] designed the first monocular depth estimation network by learning to predict coarse depth maps, which are then refined by a second network. Laina et al. [32] improved the latter model by using only convolutional layers in a single CNN. The required ground truth often limits these methods to outdoor scenarios [22]. A way of bypassing this is to use synthetic data [39]. Narrowing down the resulting domain gap can be realized [26]. MiDaS [47] generalizes better to unknown scenes by mixing data from 3D movies. To predict high-resolution depth, most methods use multi-scale features or post processing [41, 69] which complicates learning. If not trained on a massive set of data, these methods show limited generalization capabilities.

**Self-Supervision.** Self-supervised monocular methods try to circumvent this issue. The first such methods [19, 66] propose to use stereo images to train a network for depth prediction. With it, the left image is warped into the right where photometric consistency serves as training signal. Monodepth [24] added a left-right consistency loss to mutually leverage warping from one image into the other. Even though depth quality improves, it requires synchronized image pairs. Monocular training methods are developed that use only one camera where frames in a video are leveraged for the warping with simultaneously estimated poses between them. This task is more intricate, however, Monodepth2 [25] reduces the accuracy gap between the stereo and monocular training by automasking and with a minimum reprojection loss. A large body of work further improves the task [10, 33, 46, 55, 68] and investigates temporal consistency [38, 50, 64]. To compare the effect of various supervision signals for monocular depth estimation, we utilized the ResNet backbone of the popular Monodepth2 [25] together with its various training strategies.

### 2.3. Reconstruction and Novel View Synthesis

The 3D geometry of a scene can be reconstructed from 2D images and optionally their depth maps [43]. Scenes are stored explicitly or implicitly. Typical explicit representation include point clouds or meshes [11] while popular implicit representation are distance fields [71] which provide the scene as a level set of a given function, or neural fields.
where the scene is stored in the weights of a network [67]. NeRFs. Due to their photorealism in novel view synthesis, recent advances around neural radiance fields (NeRF) [42] experience severe attention. In this setup, one network is trained on a posed set of images to represent a scene. The method optimizes for the prediction of volume density and view-dependent radiance within a volume. Integration along query rays allows to synthesize novel views of static and deformable [44] scenes. Most noticeable recent advances extend the initial idea to unbounded scenes of higher quality with Mip-NeRF 360 [5] or factor the representation into low-rank components with TensoRF [9] for faster and more efficient usage. Also robustness to pose estimates and calibration are proposed [34, 63]. While the initial training was computationally expensive, methods have been developed to improve inference and training. With spherical harmonics spaced in a voxel grid structure, Plenoxels [16] speed up processes even without a neural network and interpolation techniques [59] accelerate training. Geometric priors such as sparse and dense depth maps can regularize convergence, improve quality and training time [13, 49]. Besides recent works on methods themselves, [48] propose to leverage real world objects from crowd-sourced videos on a category level to construct a dataset to evaluate novel view synthesis and category-centric 3D reconstruction methods.

We make use of most recent NeRF advances and analyse the impact of sensor-specific depth priors in [49] for the task of implicit scene reconstruction. To neglect the influence of pose estimates and produce highly accurate data, we leverage the robotic pose GT of our dataset.

3. Data Acquisition & Sensor Modalities

We set up scenes composed of multiple objects with different shapes and materials to analyse sensor characteristics. 3D models of photometrically challenging objects with reflective or transparent surfaces are recorded with high quality a priori and aligned to the scenes. Images are captured from a synchronised multi-modal custom sensor mounted at a robot end-effector to allow for precise pose camera measurements [61]. High-quality rendered depth can be extracted a posteriori from the fully annotated scenes for the viewpoint of each sensor. The acquisition pipeline is depicted in Fig. 2.

Previous 3D and depth acquisition setups [8, 12, 56] scan the scene as a whole which limits the quality by the used sensor. We instead separately scan every single object, including chairs and background, as well as small household objects a priori with two high-quality structured light object scanners. This process significantly pushes the annotation quality for the scenes as the robotic 3D labelling process only has a point RMSE error of 0.80 mm [61]. For comparison, a Kinect Azure camera induces a standard deviation of 17 mm in its working range [35]. The accuracy allows us to investigate depth errors arising from sensor noise objectively, as shown in Fig. 3, while resolving common issues of imperfect meshes in available datasets (cf. Fig. 1, left).

3.1. Sensor Setup & Hardware Description

The table-top scanner (EinScan-SP, SHINING 3D Tech. Co., Ltd., Hangzhou, China) uses a rotating table and is designed for small objects. The other is a hand-held scanner (Artec Eva, Artec 3D, Luxembourg) which we use for larger objects and the background. For objects and areas with challenging material, self-vanishing 3D scanning spray (AESUB Blue) is used. For larger texture-less areas such as tables and walls we temporarily attach small markers [20] to the surface to allow for relocalization of the 3D scanner. The robotic manipulator is a KUKA LBR iiwa 7 R800 (KUKA Roboter GmbH, Germany) with a position accuracy of ±0.1 mm. We validated this during our pivot calibration stage (Fig. 2 b) by calculating the 3D location of the tool tip (using forward kinematics and hand-tip calibration) while varying robot poses. The position varied in [−0.158, 0.125] mm in line with this. Our dataset features a unique multi-modal setup with four different cameras, which provide four types of input images (RGB, polarization, stereo, Indirect ToF (I-ToF) correlation) and three different depth images modalities (Direct ToF (D-ToF), I-ToF, Active Stereo). RGB and polarization images are acquired with a Phoenix 5.0 MP Polarization camera (PHX05051- QC, LUCID Vision Labs, Canada) equipped with a Sony PolarSens sensor (IMX264MYR CMOS, Sony, Japan). To acquire stereo images, we use an Intel RealSence D435 (In-
4. Methodology

The dataset described above allows for the first time for rigorous, in-depth analysis of different depth sensor modalities and a detailed quantitative evaluation of learning-based dense scene regression methods when trained with varying supervision signals. We focus on the popular tasks of monocular depth estimation and implicit 3D reconstruction with the application of novel view synthesis.

4.1. Depth Estimation

To train the depth estimation from a single image, we leverage the widely adopted architecture from [25]. We train an encoder-decoder network with a ResNet18 encoder and skip connections to regress dense depth. Using different supervision signals from varying depth modalities allows to study the influence and the characteristics of the 3D sensors. Additionally, we analyze whether complementary semi-supervision via information of the relative pose between monocular acquisitions and consecutive image information of the moving camera can overcome sensor issues.

We further investigate the network design influence on the prediction quality for the supervised case. For this, we train two high-capacity networks with transformer backbones on our data, namely DPT [46] and MIDAS [47].

Dense Supervision In the fully supervised setup, depth modalities from the dataset are used to supervise the prediction of the four pyramid level outputs after upsampling to the original input resolution with: \( \mathcal{L}_{\text{supervised}} = \sum_{i=1}^{4} \| \hat{D}_i - D \|_1 \), where \( D \) is the supervision signal for valid pixels of the depth map and \( \hat{D}_i \) the predicted depth at pyramid scale \( i \).

Self-Supervision Depth and relative pose prediction between consecutive frames of a moving camera can be formulated as coupled optimization problem. We follow established methods to formulate a dense image reconstruction loss through projective geometric warping [25]. In this process, a temporal image \( I_t \) at time \( t \) is projectively transformed to the frame at time \( t \) via:

\[
I_{t' \rightarrow t} = I_t \left( \text{proj}(D_t, T_{t' \rightarrow t}, K) \right),
\]

where \( D_t \) is the predicted depth for frame \( t \), \( T_{t' \rightarrow t} \) the relative camera pose, and \( K \) the camera intrinsics. The photometric reconstruction error [25, 50, 64] between image \( I_x \) and \( I_y \), given by:

\[
E_{\text{pe}}(I_x, I_y) = \alpha \frac{1 - \text{SSIM}(I_x, I_y)}{2} + (1 - \alpha) \| I_x - I_y \|_1
\]

is computed between target frame \( I_t \) and each source frame \( I_s \) with \( s \in S \). The pixel-wise minimum error is retrieved to finally define \( \mathcal{L}_{\text{photo}} \) over \( S = \{ t - F; t + F \} \) as

\[
\mathcal{L}_{\text{photo}} = \min_{s \in S} E_{\text{pe}}(I_t, I_{s \rightarrow t}).
\]

The edge-aware smoothness \( \mathcal{L}_s \) is applied [25] to encourage locally smooth depth estimations with the mean-normalized inverse depth \( \mathcal{L}_s = \| \nabla \bar{d}_t \| e^{-|\nabla d_t|} + \| \nabla \bar{d}_y \| e^{-|\nabla d_y|} \). The final training loss for the self-supervised setup is:

\[
\mathcal{L}_{\text{self-supervised}} = \mathcal{L}_{\text{photo}} + \lambda_s \cdot \mathcal{L}_s.
\]
Semi-Supervision For the semi-supervised training, the ground truth relative camera pose is leveraged. The predicted depth estimate is used to formulate the photometric image reconstruction. We also enforce the smoothness loss as detailed above.

Data Fusion Despite providing high accuracy ground truth, our annotation pipeline is time-consuming. One may ask whether this cannot be done with multi-view data aggregation. We therefore compare the quality against the dense structure from motion method Kinect Fusion [43] and an approach for TSDF Fusion [74]. The synchronized sensor availability allows also to investigate and improve sensor fusion pipelines. To illustrate the impact of high quality GT for this task, we also train the recent raw ToF+RGB fusion network Wild-ToFu [27] on our dataset.

4.2. Implicit 3D Reconstruction

Recent work on implicit 3D scene reconstruction leverages neural radiance fields (NeRF) [42]. The technique works particularly well for novel view synthesis and allows to render scene geometry or RGB views from unobserved viewpoints. Providing additional depth supervision regularizes the problem such that fewer views are required and training efficiency is increased [13, 49]. We follow the motivation of [49] and leverage different depth modalities to serve as additional depth supervision for novel view synthesis. Following NeRF literature [42, 49], we encode the radiance field for a scene in an MLP $F_\theta$ to predict colour $C = [r, g, b]$ and volume density $\sigma$ for some 3D position $x \in \mathbb{R}^3$ and viewing direction $d \in \mathbb{S}^2$. We use the positional encoding from [49]. For each pixel, a ray $r(t) = o + td$ from the camera origin $o$ is sampled through the volume at location $t_k \in [t_n, t_f]$ between near and far planes by querying $F_\theta$ to obtain colour and density:

$$C(r) = \sum_{k=1}^{K} w_k c_k$$

with $w_k = T_k (1 - \exp(-\sigma_k \delta_k))$, $T_k = \exp\left(-\sum_{k'=1}^{K} \sigma_k' \delta_k'\right)$ and $\delta_k = t_{k+1} - t_k$.

The NeRF depth $\hat{z}(r)$ is computed by: $\hat{z}(r) = \sum_{k=1}^{K} w_k t_k$ and the depth regularization for an image with rays $\mathcal{R}$ is:

$$\mathcal{L}_D = \sum_{r \in \mathcal{R}} |\hat{z}(r) - z(r)| \cdot \mathbb{1}(r)$$

where $z(r)$ is the depth of the sensor. Using the mean squared error (MSE) loss $\mathcal{L}_{\text{colour}} = \text{MSE}(C, C)$ for synthesized colours, the final training loss is:

$$\mathcal{L}_{\text{NeRF}} = \mathcal{L}_{\text{colour}} + \lambda_D \cdot \mathcal{L}_D$$

5. Sensor Impact for Dense 3D Vision Tasks

We train a series of networks for the task of monocular depth estimation and implicit scene reconstruction.

5.1. Depth Estimation

Results for monocular depth estimation with varying training signal are summarized in Table 2 and Fig. 5. We report average results for the scenes and separate performances for background, objects, and materials of different photometric complexity. The error varies from background to objects. Their varying photometric complexity can explain this. Not surprisingly, the ToF training is heavily influenced by reflective and transparent object material, where the active stereo camera can project some patterns onto diffusely reflective surfaces. Interestingly, the self- and semi-supervised setups help to recover information in these challenging setups to some extent, such that these cases even

Table 2. Depth Prediction Results for Different Training Signals. Top: Dense supervision from different depth modalities. Bottom: Evaluation of semi-supervised (pose GT) and self-supervised (mono and mono+stereo) training. The entire scene (Full), background (BG), and objects (Obj) are evaluated separately. Objects material is further split into textured, reflective and transparent. Best and 2nd best RMSE in mm are indicated.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I-ToF</td>
<td>113.29</td>
<td>111.13</td>
<td>119.72</td>
<td>54.45</td>
<td>87.84</td>
<td>207.89</td>
</tr>
<tr>
<td>D-ToF</td>
<td>77.97</td>
<td>69.87</td>
<td>112.83</td>
<td>37.88</td>
<td>71.59</td>
<td>207.82</td>
</tr>
<tr>
<td>Active Stereo</td>
<td>72.38</td>
<td>71.94</td>
<td>61.13</td>
<td>20.90</td>
<td>52.43</td>
<td>87.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sup.</th>
<th>Pose</th>
<th>154.87</th>
<th>158.67</th>
<th>65.42</th>
<th>57.22</th>
<th>37.78</th>
<th>61.86</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>180.34</td>
<td>183.65</td>
<td>85.51</td>
<td>84.26</td>
<td>48.80</td>
<td>49.62</td>
</tr>
<tr>
<td></td>
<td>M+S</td>
<td>159.80</td>
<td>161.65</td>
<td>82.16</td>
<td>71.24</td>
<td>63.92</td>
<td>66.48</td>
</tr>
</tbody>
</table>
outperform the ToF supervision for photometrically challenging objects. In contrast, simpler structures (such as the background) benefit from the ToF supervision. This indicates that sensor-specific noise is learnt and reveals that systematic errors of learning approaches cannot be evaluated if such 3D devices are used for ground truth evaluation without critical analysis. This might ultimately lead to incorrect result interpretations, particularly if self-supervised approaches are evaluated against co-modality sensor data. The table also discloses that the mutual prediction of inter-frame poses in self-supervision indoor setups is challenging, and accurate pose labels can have an immediate and significant impact on the depth results (Pose vs. M).

Fig. 6 shows that multi-view data aggregation in the form of dense SfM fails to reproduce highly reliable 3D reconstructions. In particular transparent and diffuse texture-less objects pose challenges to both Active Stereo and D-ToF. These can neither be recovered by the Kinect Fusion pipeline [43] nor by the TSDF Fusion implementation of Open3D [74] for which we use the GT camera poses. Inherent sensor artefacts are present even if depth maps from different viewpoints are combined. This quality advantage justifies our expensive annotation setup. We further analysed the results of training runs with DPT [46] and MIDAS [47], which we train from scratch. While these more complex architectures with higher capacity show the same trend and also learn sensor noise, the training time is significantly longer. More details are provided in the supplementary material. From the previous results, we have seen that ToF depth is problematic for translucent and reflective material. Fig 7 illustrates that an additional co-modal input signal at test time can cure these effects partly. It can be observed that the use of additional RGB data in [27] reduces the influence of MPI and resolves some material-induced depth artefacts. Our unique dataset also inspires cross-modal fusion pipelines’ development and objective analysis.

5.2. Implicit 3D Reconstruction & View Synthesis

Our implicit 3D reconstruction generates novel views for depth, normals and RGB with varying quality. If trained with only colour information, the NeRF produces convincing RGB views with the highest PSNR (cf. Fig. 8 and Table 3). However, the 3D scene geometry is not well reconstructed. In line with the literature [13, 49], depth regularization improves this (e.g. on texture-less regions). Regularising with different depth modalities makes the sensor noise of I-ToF, AS, and D-ToF clearly visible. While the RMSE behaves similarly to the monocular depth prediction results with AS as best, followed by D-ToF and I-ToF. The cosine similarity for surface normal estimates confirms this trend. The overall depth and normal reconstruction for AS are very noisy, but depth error metrics are more sensitive for significant erroneous estimates for reflective and translucent objects. Prior artefacts of the respective sensor influence the NeRF and translate into incorrect reconstructions (e.g. errors from D-ToF and I-ToF for translucent material or noisy background and inaccurate depth discontinuities at edges for AS). Interestingly, the D-ToF prior can improve
The results of an implicit scene reconstruction with a Neural Radiance Field (NeRF) are shown. Images are synthesised for depth, surface normals and RGB for an unseen view, which is shown together with the prediction errors. The columns allow us to compare different methods where a NeRF [42] is trained solely on RGB (first column) and various depth maps for regularisation as proposed in [49]. The last column illustrates synthesised results from training with GT depth for comparison. Differences are visible, especially for the partly reflective table edges, the translucent bottle and around depth discontinuities.

the overall reconstruction for most of the scene but fails for the bottle, where the AS can give better depth priors. This is also visible in the synthesised depth. Leveraging synthetic depth GT (last row) mitigates these issues and positively affects the view synthesis with higher SSIM.

6. Discussion & Conclusion

This paper shows that questioning and investigating commonly used 3D sensors helps to understand their impact on dense 3D vision tasks. For the first time, we make it possible to study how sensor characteristics influence learning in these areas objectively. We quantify the effect of various photometric challenges, such as translucency and reflectivity for depth estimation, reconstruction and novel view synthesis and provide a unique dataset to stimulate research in this direction. While obvious sensor noise is not "surprising", our dataset quantifies this impact for the first time. For instance, interestingly, D-ToF supervision is significantly better suited (13.02 mm) for textured objects than AS, which in return surpasses I-ToF by 3.55 mm RMSE (cf. 2). Same trend holds true on mostly texture-less backgrounds where D-ToF is 37% more accurate than I-ToF. For targeted analysis and research of dense methods for reflective and transparent objects, a quantitative evaluation is of utmost interest - while our quantifiable error maps allow specifying the detailed deviations. Although our dataset tries to provide scenes with varying backgrounds, the possible location of the scene is restricted due to the limited working range of the robot manipulator. Aside from our investigations and the evaluation of sensor signals for standard 3D vision tasks, we firmly believe that our dataset can also pave the way for further investigation of cross-modal fusion pipelines.

Table 3. Novel View Synthesis from Implicit 3D Reconstruction. Evaluation against GT for RGB, depth and surface normal estimates for different optimisation strategies (RGB-only for supervision and + respective sensor depth). We indicate best, 2nd best and 3rd best. Depth metrics in mm.
References


[71] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch:
Learning local geometric descriptors from rgb-d reconstructions. In CVPR, 2017. 3


