A Plenoptic 3D Vision System - Supplement

AGASTYA KALRA, Intrinsic Innovation LLC, USA
VAGE TAAMAZYAN, Intrinsic Innovation LLC, USA
ALBERTO DALL'OLIO, Intrinsic Switzerland GmbH, Switzerland
RAGHAV KHANNA, Intrinsic Switzerland GmbH, Switzerland
TOMAS GERLICH, Intrinsic Innovation LLC, USA
GEORGIA GIANNOPOLOU, Intrinsic Switzerland GmbH, Switzerland
GUY STOPPI, Intrinsic Canada Corporation, Canada
DANIEL BAXTER, Intrinsic Innovation LLC, USA
ABHIJIT GHOSH, Intrinsic Innovation LLC, USA
RICK SZELISKI, Google DeepMind, USA
KARTIK VENKATARAMAN, Intrinsic Innovation LLC, USA

In this supplement we provide additional details of experiments performed in the main paper.

ACM Reference Format:
Agastya Kalra, Vage Taamazyan, Alberto Dall'olio, Raghav Khanna, Tomas Gerlich, Georgia Giannopolou, Guy Stoppi, Daniel Baxter, Abhijit Ghosh, Rick Szeliski, and Kartik Venkataraman. 2024. A Plenoptic 3D Vision System - Supplement. *ACM Trans. Graph.* 1, 1 (September 2024), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 AUTO CALIBRATION

We describe each step of our automatic calibration procedure below.

(1) **Pattern Projection**
A pseudo-random dot pattern is projected onto a surface visible to all cameras in the system. To ensure visibility in high ambient light environments, the pattern is projected at a wavelength of 940nm in the near-infrared spectrum.

(2) **Image Capture**
Each IR camera captures two images, one with the dots projected and one with the projections off.

(3) **Dot Centroid Localization and Correspondence**
- Dot difference images are computed by subtracting the background images from the projector-on images and then blurring with a width 3 Gaussian to reduce noise.
- Dot centroid candidates are identified by local maxima detection within a sliding 3×3 window.
- Sub-pixel refinement of dot centroids is achieved by fitting a quadratic function to a 5x5 neighborhood around each peak.
- The strongest *1000* peaks per image are selected for subsequent correspondence establishment.
- A coarse camera-to-camera transform is estimated using the Open3D global registration pipeline using FPFH features [Rusu et al. 2009] along with the RANSAC algorithm [Zhou et al. 2018], then refined with the Iterative Closest Point (ICP) algorithm [Besl and McKay 1992].
- Dot correspondences between cameras are established for each projector active image set by projecting the 3D points associated with detected dots from one camera to another using the coarse transform and finding the nearest corresponding detected dot within a specified threshold (k=2 pixels).

(4) **Extrinsic Parameter Estimation**
The established dot correspondences are used to refine the camera-to-camera transform using ICP, aligning the 3D points associated with corresponding dots across cameras.

(5) **Bundle Adjustment Optimization**
The initial estimates can be further refined using bundle adjustment [Triggs et al. 2000] to minimize the re-projection error of the corresponding dots in all infra-red camera images across all units.

(6) **Dot Correspondence Refinement (Optional)**
Lucas-Kanade optical flow [Lucas and Kanade 1981] is employed to refine the sub-pixel accuracy of dot correspondences after Step 3.

Authors' addresses: Agastya Kalra, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, agastyak@intrinsic.ai; Vage Taamazyan, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, vage@intrinsic.ai; Alberto Dall'olio, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, dallolio@intrinsic.ai; Raghav Khanna, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, raghavkhanna@intrinsic.ai; Tomas Gerlich, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, tgerlich@intrinsic.ai; Georgia Giannopolou, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, giannopolou@intrinsic.ai; Guy Stoppi, Intrinsic Canada Corporation, 111 Richmond St W, Toronto, M5H2G4, Canada, guystoppi@intrinsic.ai; Daniel Baxter, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, danielbaxter@intrinsic.ai; Abhijit Ghosh, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, ghoshabhijit@intrinsic.ai; Rick Szeliski, Google DeepMind, 100 Mayfield Ave., Mountain View, 94043, USA, szeliski@google.com; Kartik Venkataraman, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, kartikvp@intrinsic.ai.

0730-0301/2024/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

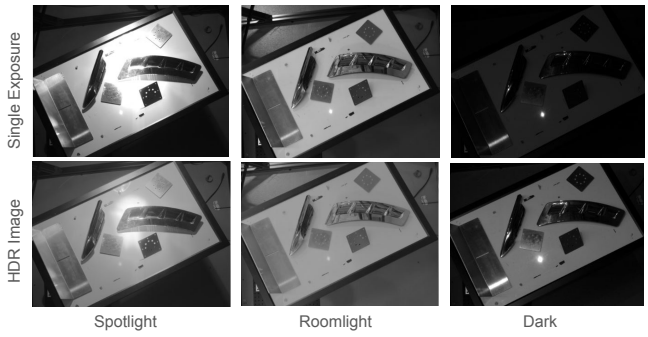

Fig. 1. **Our testing dataset collects each scene in 3 different lighting conditions**. Our system captures multiple exposures and fuses them using HDR to help alleviate the challenge of multiple lighting conditions.

(7) **Extrinsic Parameter Validation**
The calibration procedure can be repeated to validate existing calibration parameters in a production setting, ensuring that the difference between the existing and newly estimated parameters is within a desired tolerance.

2 LIGHTING

In Figure 1, we show that each scene was captured with different lighting conditions, and that HDR allows us to get comparable images across these lighting conditions.

3 DPSNET

Figure 6 provides a visual comparison, demonstrating the strong generalization of our model to DPSNet's polarization data, despite being trained solely on synthetic data.

4 CRESTEREO BASELINE

We evaluated the performance of CREStereo [Li et al. 2022], a state-of-the-art stereo matching model, on our dataset (Table 1) using the publicly available weights. However, we observed significant challenges in accurately reconstructing industrial objects and intricate bin walls. Due to CREStereo's lack of multi-modal fusion capabilities, we restricted our comparison to RGB-only inputs. Notably, even with this constraint, our model, trained solely on our synthetic RGB data, surpassed the performance of CREStereo trained exclusively on RGB. Moreover, as demonstrated in the main paper, incorporating additional modalities further enhances performance, underscoring the potential benefits of a multi-modal approach for 3D reconstruction in complex industrial environments.

5 TRAINING DATA CREATION

This section details our synthetic data generation pipeline and polarized data augmentations.

5.1 Physically Accurate Data Rendering

As mentioned in the main paper, our P-Stereo network training relies entirely on synthetic data. Generating physically accurate polarization signals is challenging, and to our knowledge, Mitsuba 3 is the only physics-based renderer currently capable of it [Jakob et al. 2022]. Physics-based rendering requires accurate simulation of all material properties and their spectral dependence, global illumination, and light source properties. Mitsuba 3 can handle all this and simulate accurate Stokes vectors, which are then converted into AOLP and DOLP.

5.1.1 Object Placement. We created a dataset of around 1,000 diverse CAD models representing various abstract shapes and real objects. During each scene generation, a random set of CAD models is sampled, followed by a random selection of one of several placement strategies: random placement in 3D space, gravity-based placement, parallel to baseline placement (aligning parts with the baseline, often making stereo reconstruction more challenging for parts with long rims or textureless regions). Placement is always within a working volume of 500 to 5,000 mm. Each part can be randomly scaled during placement.

5.1.2 Materials. Each part is randomly assigned one of the predefined materials from the Mitsuba 3 library. Materials include dielectrics (including transparent ones) and various metals. We only select materials supporting realistic physics-based rendering with polarization, excluding the principled BSDF and other materials not directly related to physics properties. Some randomly selected objects are covered with textures from a library of over 3,000 different textures.

5.1.3 Lighting. We randomly set point or shaped-based light sources in the scene. While Mitsuba 3 doesn't allow directly specifying light source polarization properties, we randomly generate linear polarizers in front of some light sources to ensure polarized incident light. Additionally, we randomly place various objects above the cameras to introduce polarization generated by interreflections. Mitsuba 3 inherently accounts for how interreflections alter light polarization, enhancing rendering realism. We also randomly set environmental lighting from a selection of over 1,300 different environment maps.

5.1.4 IR Dots. We generate a semi-random IR dot pattern and use a projector light source in Mitsuba 3 to project it onto the scene. Mitsuba 3 accounts for object reflectivity, with more reflective objects returning little or no IR dot signal, similar to real-life scenarios. We also ensure virtual IR cameras are shifted against RGB cameras, mimicking real units.

5.1.5 Polarization. We use the Stokes vector simulated by Mitsuba to compute AOLP and DOLP aligned with the RGB camera. Currently, we don't simulate separate P60 and P120 cameras in this dataset, but we plan to do so in the future for even closer simulation of stereo units. Material and lighting randomization ensures sufficiently randomized polarization signals rendered at the camera.

5.2 Polarized Data Augmentation

Our polarized data augmentation strategy aims to simulate the physical properties of polarization while introducing noise. This approach is guided by several key observations:

Experiment	Input Data	Units	Roomlight		Spotlight		Dark		Average	
			FNR, %	FPR, %	FNR, %	FPR, %	FNR, %	FPR, %	FNR, %	FPR, %
CREStereo	RGB	1	42.1	1.0	41.8	0.9	43.6	1.1	42.5	1.0
Ours	RGB	1	13.2	2.1	13.3	**2.1**	17.4	3.2	14.6	2.5
Ours (Final)	RGB + IR + Polar	2	**8.6**	**1.8**	**8.1**	**2.1**	**9.6**	**2.1**	**8.7**	**2.0**

Table 1. **Our model trained with our synthetic data outperforms a CREStereo trained on many existing public datasets.** This shows that our dataset is challenging and cannot easily be solved by existing state-of-the-art stereo networks. Furhtermore, our Plenoptic Stereo Architecture shows large improvements on our dataset, showing the benefit of multiple baselines and multiple modalities.

- **AOLP-DOLP Correlation:** The quality of the Angle of Linear Polarization (AOLP) signal is strongly linked to the Degree of Linear Polarization (DOLP). Weak DOLP signals often correspond to unreliable AOLP measurements.
- **AOLP Variability:** The AOLP, measured in degrees, is highly sensitive to lighting conditions and viewpoint. Our goal is to encourage the network to learn the underlying AOLP texture rather than specific values.
- **DOLP-Dependent Noise:** The DOLP noise level is signal dependent with higher DOLP values exhibiting more noise.

To replicate these phenomena, we propose the following data augmentation pipeline:

(1) **Random Blurring:** To simulate focus imperfections, we apply Gaussian blurring to both the AOLP and DOLP images with a 50% probability.
(2) **AOLP Offsets:** We introduce independent rotations of up to 30° to the AOLP of the left and right cameras. This discourages the model from expecting identical AOLP values across viewpoints.
(3) **AOLP Noise:** Gaussian noise is applied independently to the AOLP image, scaled by the DOLP. This means that when DOLP is 0, the AOLP receives 180° Gaussian noise. For DOLP values above 0.2, the maximum AOLP noise is capped at 10°. We find this approach produces a more realistic noise profile.
(4) **AOLP Renormalization:** To maintain the AOLP within its valid range of $[0, \pi]$, we apply a modulo operation with π.
(5) **DOLP Scaling:** The DOLP is randomly scaled by up to 30%.
(6) **DOLP Noise:** Gaussian noise is added to the DOLP, with a standard deviation equal to 15% of the DOLP value.

Visual examples of synthetic scenes generated using this augmentation pipeline are presented in Figure 5. We also provide the raw synthetic data for the read in Figure 4. During training we also perform cropping, scaling, and standard RGB photometric augmentations.

6 COLLISION AVOIDANCE METRIC ABLATION

In the main paper, we employ the Collision Avoidance Metric [Taamazyan et al. 2024] with a collision Z threshold of 10mm. All other parameters are consistent with [Taamazyan et al. 2024]. Here, we demonstrate that the choice of Z threshold does not alter ordering of methods presented in the main paper. Metrics were calculated, averaged across all scenes, for tolerances ranging from 5mm to 20mm in Figures 2 and 3.

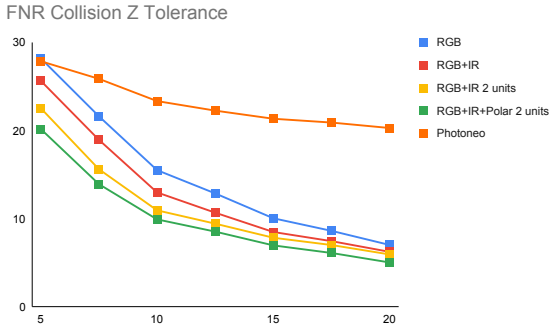

Fig. 2. FNR for collision Z thresholds from 5mm to 20mm. Our full system outperforms others across all thresholds.

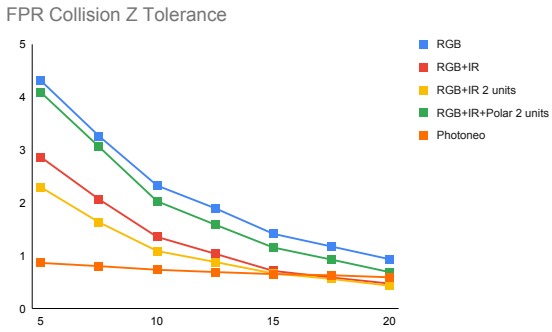

Fig. 3. FPR for collision Z thresholds from 5mm to 20mm. All methods converge to similar performance around 20mm.

In Figure 2, we observe that the advantage of multiple units is higher at lower collision thresholds. This verifies that adding additional units improves our triangulation accuracy. We also show that the FNR curve for our full system (2 units, all modalities) consistently outperforms other configurations across all thresholds.

In Figure 3, we see the FPR curve for our full system is slightly higher than others (except RGB), converging towards them at a 20mm threshold. We also find the structured light performance does not improve significantly with increase the threshold - likely because the system is more accurate than 5mm. These findings indicate that the choice of the Z threshold does not impact the overall claims and conclusions of the paper.

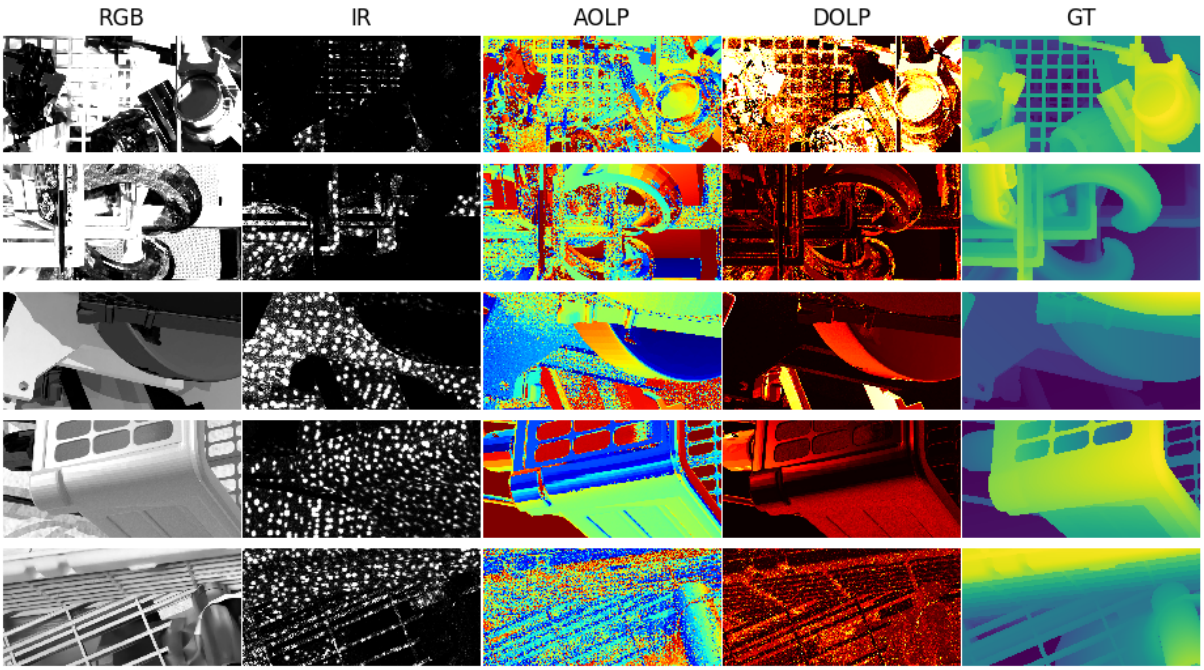

Fig. 4. **Our Synthetic Training Dataset without data augmentations**

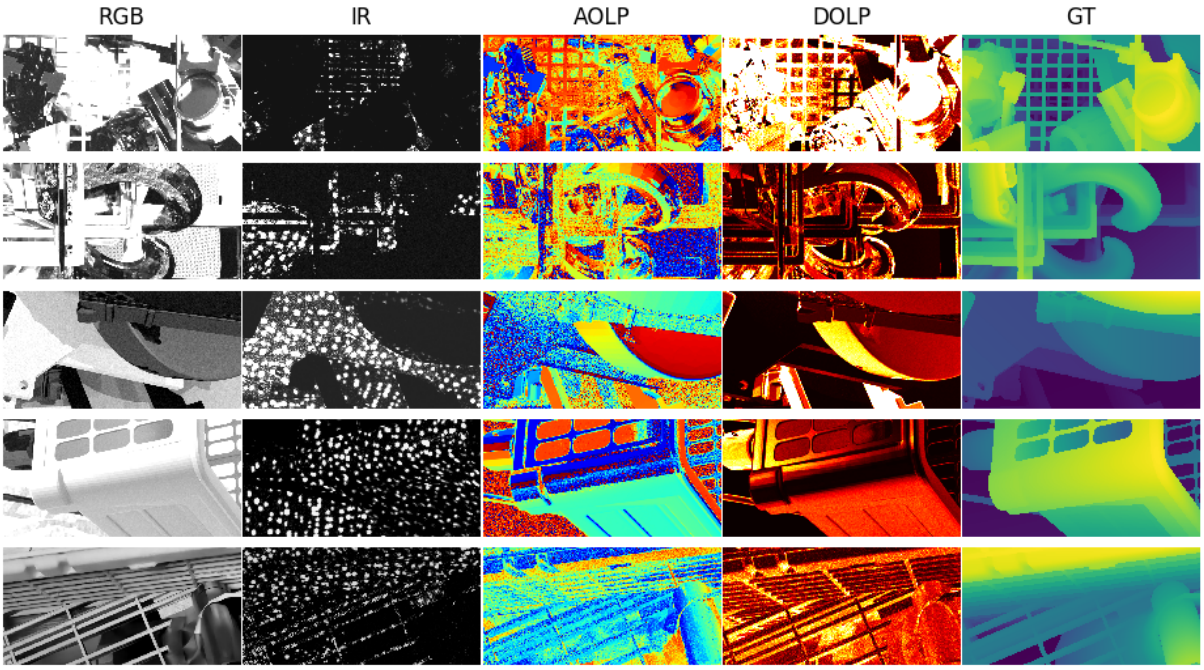

Fig. 5. **Our Synthetic Training Dataset with photometric data augmentations**

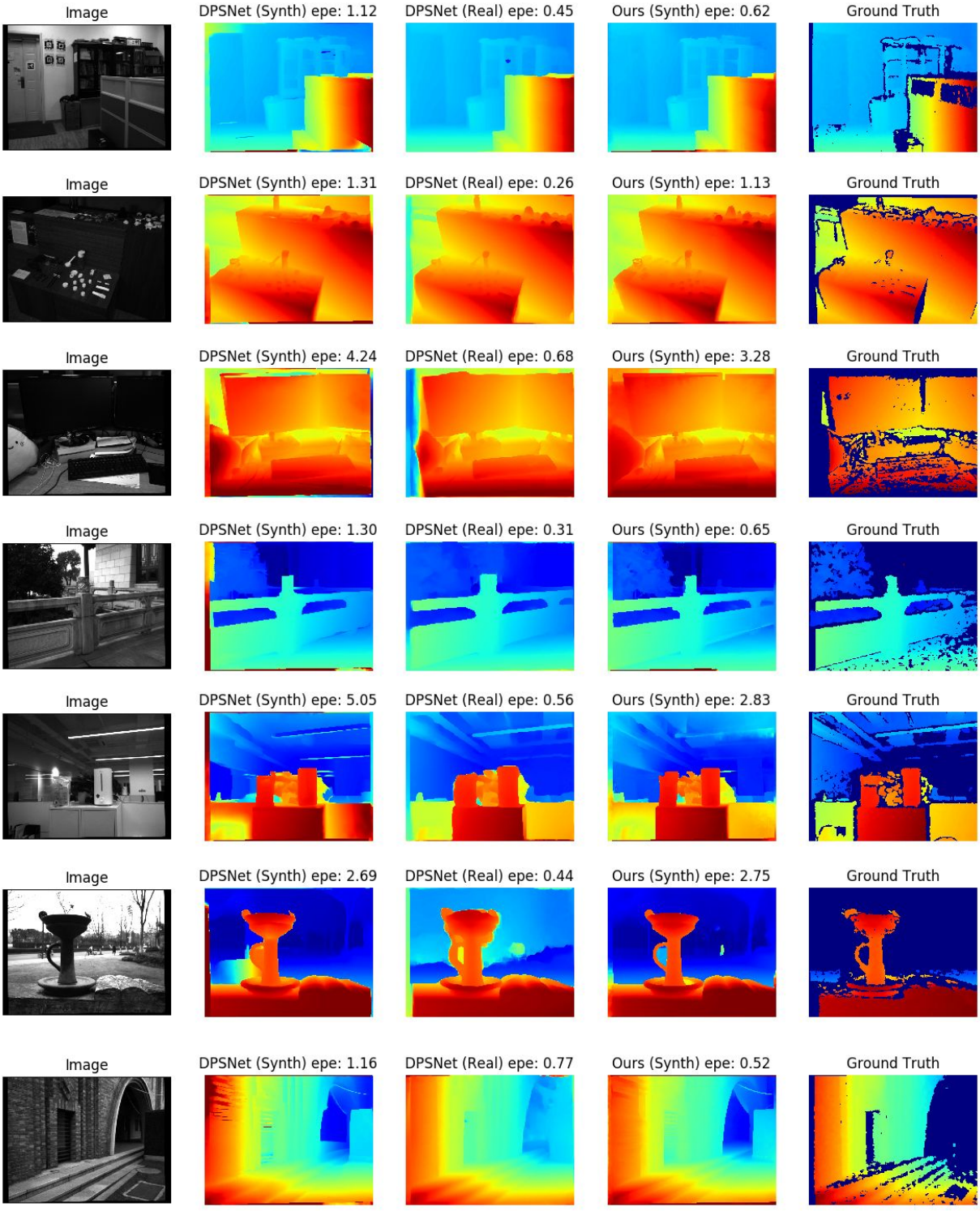

Fig. 6. **Visual comparisons of our predictions and that of DPSNet. We see that our system tends to produce more realistic estimates (see last few rows) even though it was trained only on synthetic data. It even has cleaner edges than the ground truth.**

REFERENCES

Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611. Spie, 586–606.

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. 2022. *Mitsuba 3 renderer*. https://mitsuba-renderer.org.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. 2022. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. arXiv:cs.CV/2203.11483

Bruce D Lucas and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vol. 2. 674–679.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. 2009. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation*. IEEE, 3212–3217.

Vage Taamazyan, Alberto Dall'olio, and Agastya Kalra. 2024. Collision Avoidance Metric for 3D Camera Evaluation. arXiv:cs.CV/2405.09755

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. 2000. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*. Springer, 298–372.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847* (2018).