Towards 3D Vision with Low-Cost Single-Photon Cameras

Supplementary Material

In this supplement, we describe (1) The implementation details of our reconstruction pipeline (Section A); (2) a sensitivity analysis to understand the robustness of our method w.r.t. imaging parameters (Section B); (3) A discussion of the difficulty of point cloud surface reconstruction (Section C); (4) additional qualitative results on simulated data and real world captures (Section D); and (5) further discussion of our work (Section E). We also include three short videos, which provide animated 360° views of qualitative reconstruction results. The videos can also be viewed on the project web page: https://cpsiff.github.io/towards_3d_vision/.

For sections, figures and equations, we use numbers (*e.g.*, Sec. 1) to refer to the main paper and capital letters (*e.g.*, Sec. A) to refer to this supplement.

A. Implementation Details

For both simulated and real world experiments, we use an 8-layer MLP with 256 hidden units as our SDF, f_θ, and initialize it as a sphere, centered at the origin with radius 0.3m, using geometric initialization [56]. For each transient, we sample 256 rays $\boldsymbol{\omega}$ over Ω and sample 256 points per ray. We set $\lambda_{Eikonal}$ to 0.1 across all experiments and set λ_{TV} to 0 and 0.01 respectively for the simulated and real-world experiments. We train f_θ for 300K steps using Adam [60] with a mini-batch size 2, a learning rate 0.0005, and cosine decay. The learned SDFs are converted to meshes using Marching Cubes [61].

B. Sensitivity Analysis

We perform extensive experiments to understand the robustness of our method in comparison to baselines under varying sensor parameters in simulation. All experiments are based on the Bunny scene and the parameters are varied one at a time while other parameters remain fixed at the base condition (as described in Section 4 of the main paper). To ensure strong baselines for every sensor configuration, we calibrate the thresholds t_p and t_s for the projection (threshold) and space carving baselines respectively per reconstruction. We perform a brute force search over possible thresholds and report the best Chamfer achieved. As this amounts to calibrating on the test set, the numbers reported represent the best possible performance of the baseline methods on the given data. The results of this sensitivity analysis are presented in Figure A. In what follows we discuss some of the main findings.

Sensor Placement. We study two key parameters that control sensor placement: the number of views and the minimum elevation angle at which the sensors are placed. Our method consistently outperforms all baselines in Chamfer distance by an order of magnitude across a broad range of parameter choices. In particular, our method readily supports as few as 128 views above a considerably large elevation angle of 30° without harming reconstruction quality. This robust gain in performance confirms that our method takes advantage of broad-band signal in transients not exploited by the baseline methods.

Temporal Resolution. Our system takes advantage of the temporal information in transient histograms, and therefore benefits when that information is present at a high resolution. Because of this, our method outperforms baselines by a very wide margin at a small bin size, but the margin vanishes as bins become wider than $2\mathrm{cm}$ (equivalently $66\mathrm{ps}$), because decomposing the temporal signal becomes impractical beyond this limit. Fortunately, today's commodity SPADs operate at a smaller bin size ($\sim 40\mathrm{ps}$). Baseline methods show no performance gain at small bin sizes, as they do not take advantage of the temporal resolution.

Angular Resolution. Our system resolves spatial resolution from wide-FoV sensors by taking advantage of the time dimension. In this regime, the optimal sensor field-of-view size is not obvious: a smaller FoV means more highly constrained geometry, as each histogram images a smaller region, but too small of a field-of-view means a lack of coverage and under-constrained geometry. We find that an angular resolution in the 30° to 60° range is optimal for reconstructing 3D geometry with our method on the bunny scene. Reprojection based methods benefit more from a smaller field-of-view, while space carving performs best with a wider field-of-view so that space is sufficiently carved away. In every case, our method outperforms baselines by a wide margin.

Signal-to-noise ratio (SNR). We consider three parameters that jointly impact SNR: illumination power, ambient flux, and number of illumination cycles. Our method again outperforms all baselines by a significant margin across all test conditions. Notably, the baselines fail or perform considerably worse under high ambient flux, as signal photons are blocked by background photons due to pile-up. By contrast, our method is robust against a broad range of ambient flux levels, as we model the effects of ambient flux directly.

C. Point Cloud Surface Reconstruction

We do not apply surface reconstruction to the point clouds computed by reprojection because off-the-shelf reconstruc-

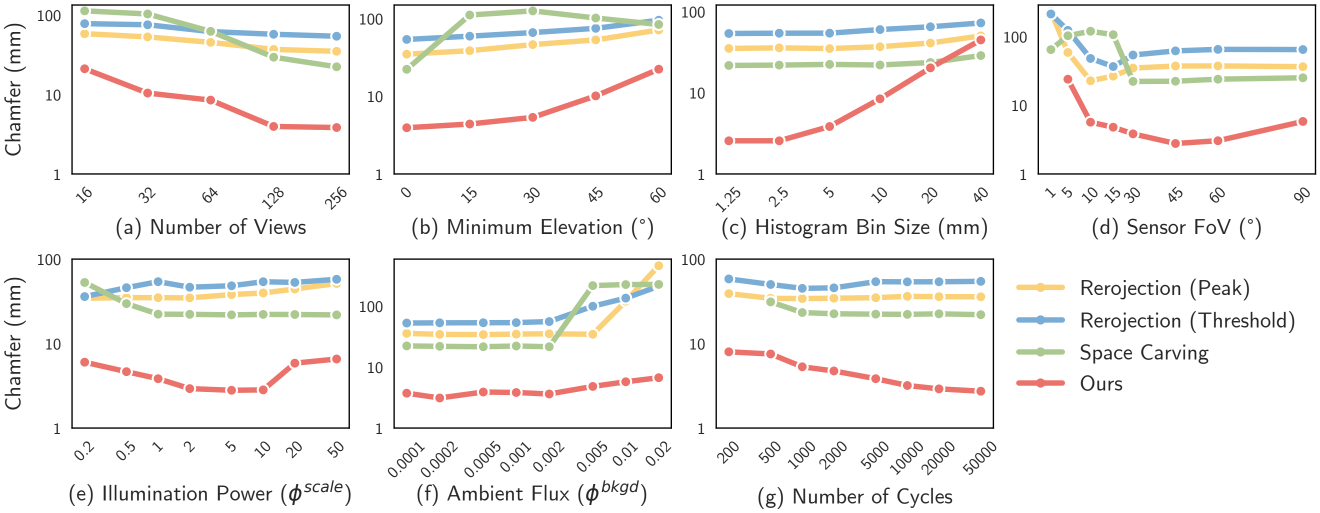

Figure A. **Sensitivity analysis of our method compared to baselines across a range of imaging parameters.** In almost every case, our method outperforms baseline methods on Chamfer distance. Missing datapoint in (d) indicates that our method failed to converge. Illumination power (e) is unit-less as it also absorbs factors like quantum efficiency and does not map directly to any real world parameter.

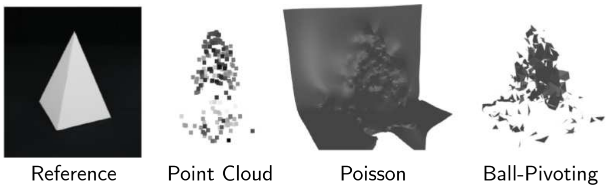

Figure B. Off-the-shelf algorithms for surface reconstruction (Poisson Surface Reconstruction [59] and The Ball-Pivoting Algorithm [57]) do not perform well on point clouds generated by reprojection.

tion techniques do not perform reliably on the generated point clouds, as shown in Figure B. Additionally, calculating Chamfer distance to the computed point cloud ensures that we are capturing the efficacy of reprojection rather than a given surface reconstruction method. We include one-way reconstruction to ground truth Chamfer distance in Table 2 to provide a metric which does not penalize the sparsity of point clouds produced by reprojection.

D. Additional Qualitative Results

More reconstruction results. We present additional qualitative results on simulated data (Figure D) and real-world captures (Figure C). These results were omitted from the main paper due to lack of space.

Surface normal visualization. Moving beyond the 3D shapes, we further examine the surface normal of our reconstructed 3D objects. The surface normal of a point $\mathbf{x}$ on the reconstructed mesh is estimated as

$$\tilde{\mathbf{n}}_{\mathbf{x}} = \frac{\nabla_{\mathbf{x}}(f_\theta(\mathrm{PE}(\mathbf{x})))}{\|\nabla_{\mathbf{x}}(f_\theta(\mathrm{PE}(\mathbf{x})))\|}, \tag{A}$$

where f_θ is the learned SDF and PE denotes the positional encoding function. The error $e_{\mathbf{x}}$ w.r.t. the ground-truth normal $\mathbf{n}_{\mathbf{x}}$ is given by

$$e_{\mathbf{x}} = |\langle \mathbf{n}_{\mathbf{x}}, \tilde{\mathbf{n}}_{\mathbf{x}} \rangle|. \tag{B}$$

Surface normal results. We provide visualizations of surface normals for simulated data in Figure E. Our method can successfully recover smoothly varying normals. Error typically occurs at edges and depth discontinuities with fast-changing normals. We hypothesize that sensors with higher temporal and spatial resolution are needed for more accurate surface normal reconstruction.

E. Further Discussion

Beyond Lambertian objects. Our method assumes a spatially uniform Lambertian BRDF, but in practice can effectively reconstruct objects with spatially varying albedo and slightly glossy appearance (*e.g.* the spray bottle). In theory, our method can easily be adapted to incorporate a parametric lighting model. Recovery of the parameters of such a model are likely possible because, by sharing information among many observations, the BRDF is effectively sampled at many incident and exitant angles. An intriguing direction for future work is investigating which BRDF parameterizations can be recovered with our imaging setup, and the effect of the reflectance model on reconstruction quality. We suspect that a non-parametric NeRF-like BRDF would not be suitable as it does not sufficiently constrain the optimization. A parametric lighting model, *e.g.* Phong [58] or Oren-Nayar [64] may appropriately constrain the optimization while allowing the model to learn a more accurate scene representation.

Runtime efficiency. Our method takes on the order of hours to reconstruct a scene, making it unsuitable for real-time applications in its current state. Future work should investigate ways to speed up forward rendering and model training. Improved importance sampling would likely yield modest improvements in convergence time. Another option is to render only summary statistics of the histogram (*e.g*. mean, peak locations or widths) rather than the entire histogram, which would likely be faster to render at the expense of yielding a lower-quality reconstruction.

Sensor pose. In this work, we used an industrial robot arm to gather posed sensor measurements. We chose this modality as it is guaranteed to provide highly accurate sensor poses, and allows control over precise sensor placement. For applications like wearable computing and distributed sensing for robotics, camera poses might be pre-calibrated and remain fixed relative to each other during operation. Alternatively, the low-cost single-photon camera could be combined with a sensor-based localization system (*e.g*., an IMU based [66] or a camera based [63] system) to recover camera pose, a setup which is standard in related works [62, 65]. Such a capture setup would allow capture of more organic and large scale scenes, which more closely mimic the potential use cases of the sensor (*e.g*. on mobile robots and drones).

Comparison to other 3D imaging modalities. Our work provides a low-cost 3D imaging system using single-photon cameras. We provide detailed comparisons between our method and baseline methods, but do not compare our reconstructions to those gathered from other 3D modalities, such as continuous wave time-of-flight [3] or LiDAR [21]. Future work should provide a comparison to these other modalities to provide insights into the niche (in terms of accuracy, size, power, etc.) filled by each.

Comparison to a NeRF with a miniature RGB camera. Using multiple calibrated RGB cameras, NeRF-like approaches can leverage photometric cues (*e.g*. correspondence and shading) to recover 3D shapes from a distributed set of cameras. *Conceptually*, NeRF-like approaches fall short under non-ideal lighting (*e.g*. low-light) or insufficient correspondence (*e.g*. textureless objects), while our method with active lighting and depth remains effective. With sufficient light and distinct texture patterns, NeRF-like approaches on RGB images will yield a higher quality reconstruction than our method, due to significantly increased data rates. Without compression, a VGA-resolution image with 8-bit color channels contains 7Mbits of information, while an image (*i.e*., transient histogram) from our sensor contains 3Kbits of information. With 2400× the information per-view, it is not surprising that NeRF with RGB cameras could outperform our method given the same number of input views. *Practically*, existing low-cost RGB cameras are larger and less power efficient than the SPAD sensors that we utilize. While miniature RGB cameras do exist (*e.g*., those used for endoscopy), they are >20× the cost of a SPAD.

Commodity sensors. One challenge for future work is a lack of hardware support for measurement and use of transient histograms. Very few low-cost sensors allow access to transient histograms, and those that do often perform preprocessing that is proprietary or undocumented. Additionally, most sensors are equipped with very low-bandwidth I^{2}C interfaces, limiting their effective FPS. We hope that manufacturers will see value in users having access to transient histogram data and support the use of this data with documentation, low-level access, and high-bandwidth interfaces in the future.

Ethical concerns. Our work presents a new method for imaging 3D objects with low-cost single-photon cameras. We do not anticipate major ethical concerns.

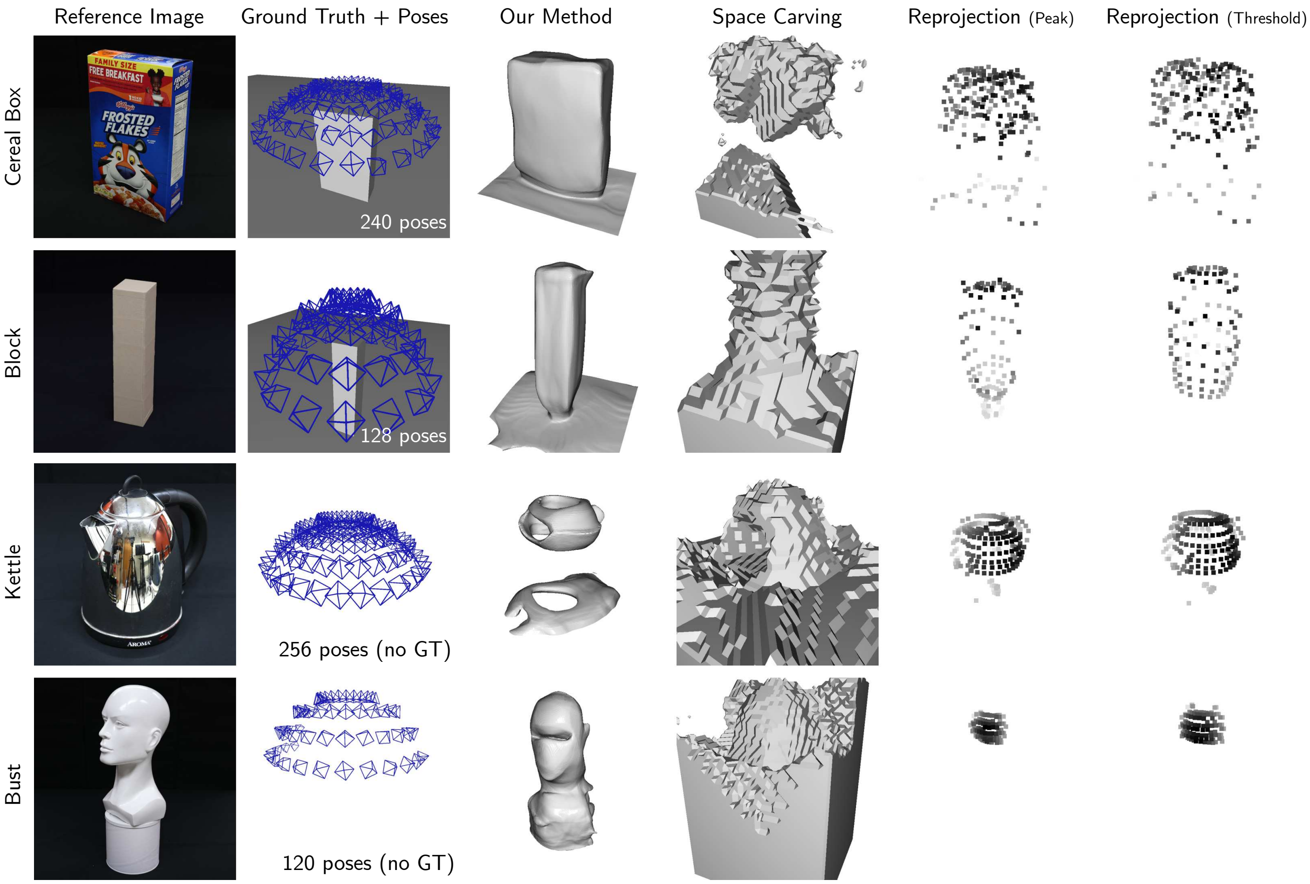

Figure C. **Additional qualitative results on real-world captures.** Our method achieves the highest reconstruction quality. Poses in column two are subsampled by a factor of two for clarity.

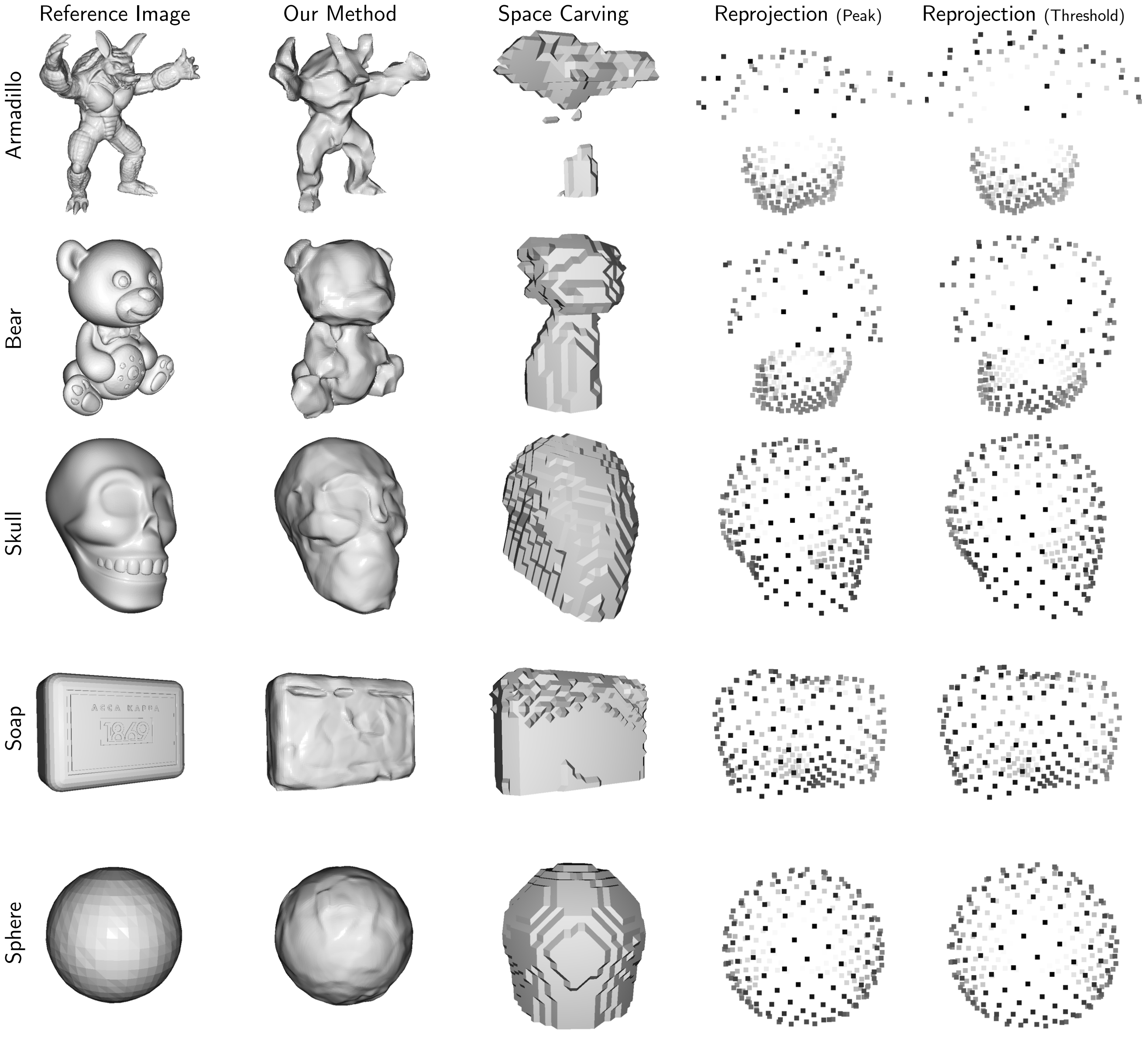

Figure D. **Additional qualitative results on simulated data.** Our method achieves the highest reconstruction quality. Space carving captures an envelope of the shape, and may carve away occupied areas in concave shapes (*e.g*. Armadillo). Reprojection gives a sparse reconstruction of convex shapes, (*e.g*. skull, soap, sphere), the scale of which may be distorted due to biases introduced by the wide field-of-view of the sensor.

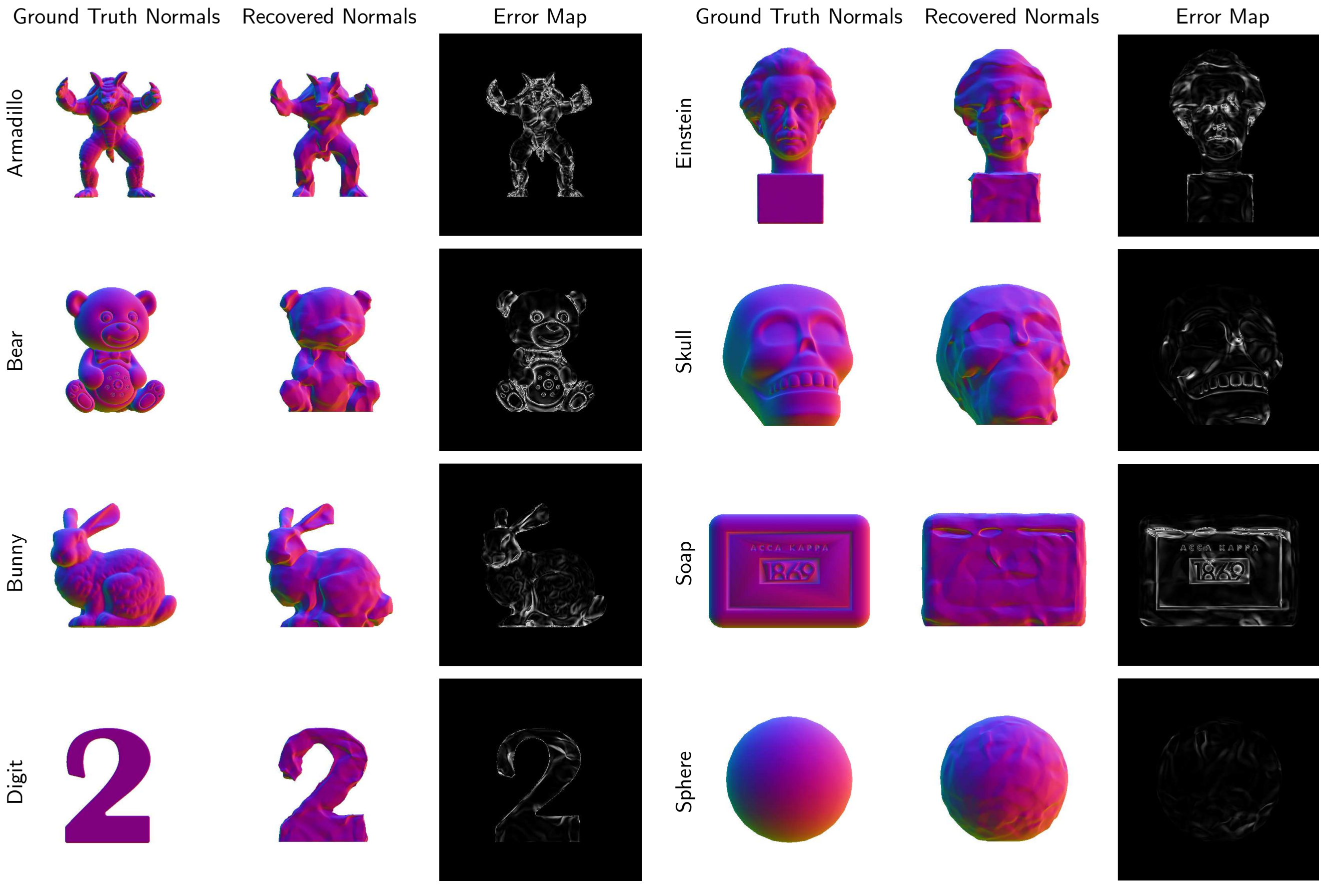

Figure E. **Visualizations of surface normals for simulated data.** Our method correctly estimates surface normals in flat regions. Error mainly occurs at edges and depth discontinuities. We hypothesize that sensors with higher temporal and spatial resolution are needed to detect rapid changes in surface normals.

Supplementary References

[56] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *CVPR*, 2020.

[57] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 5(4):349–359, 1999.

[58] Phong Bui-Tuong. Illumination for computer generated pictures. *CACM*, 1975.

[59] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006.

[60] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[61] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, 1987.

[62] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[63] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 2015.

[64] Michael Oren and Shree K Nayar. Generalization of lambert's reflectance model. In *SIGGRAPH*, 1994.

[65] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. In *RSS*, 2022.

[66] Jingang Yi, Junjie Zhang, Dezhen Song, and Suhada Jayasuriya. Imu-based localization and slip estimation for skid-steered mobile robots. In *IROS*, 2007.