The Implicit Values of A Good Hand Shake: Handheld Multi-Frame Neural Depth Refinement Supplemental Document

Ilya Chugunov[1†] Yuxuan Zhang[1] Zhihao Xia[2] Xuaner Zhang[2] Jiawen Chen[2] Felix Heide[1]

[1]Princeton University [2]Adobe

In this Supplemental Document, we present supporting material including additional results, ablation experiments, and discussion that could not otherwise fit within the main text. Specifically we provide:

- Analysis of hand shake data:
  1. Visualization of individuals' hand shake data and commentary on inter-person variance. (Section 1.1)
  2. Hand shake statistics and probabilistic view of bundle recording. (Section 1.2)
- Additional results and visualizations:
  1. Reconstruction results for additional test scenes. (Section 2.1)
  2. Photometric error maps. (Section 2.2)
  3. Reconstruction results for room-scale (>1m) scenes. (Section 2.3)
  4. Reconstruction results for close-range (<10cm) unfocused scenes. (Section 2.4)
  5. Reconstruction results for mixed-range (<0.5m + >1m) scenes. (Section 2.5)
  6. Reconstruction results for variable levels of hand shake. (Section 2.6)
  7. Comparison to RGB-guided upsampling. (Section 2.7)
- Additional ablation results:
  1. Effects of changes in sampled patch size K. (Section 3.1)
  2. Effects of changes in geometric regularization weight α. (Section 3.2)
  3. Effects of changes in number of encoding functions L. (Section 3.3)
  4. Effects of Gaussian versus square patch weighting. (Section 3.4)

Project Page: https://light.princeton.edu/hndr/

1. Additional Hand Shake Analysis

1.1. Hand Shake Visualization

In Figure 1 we present per-individual point clouds which we create by aggregating all the hand shake paths recorded by each of our 10 volunteers, as outlined in Section 4 of the main text.

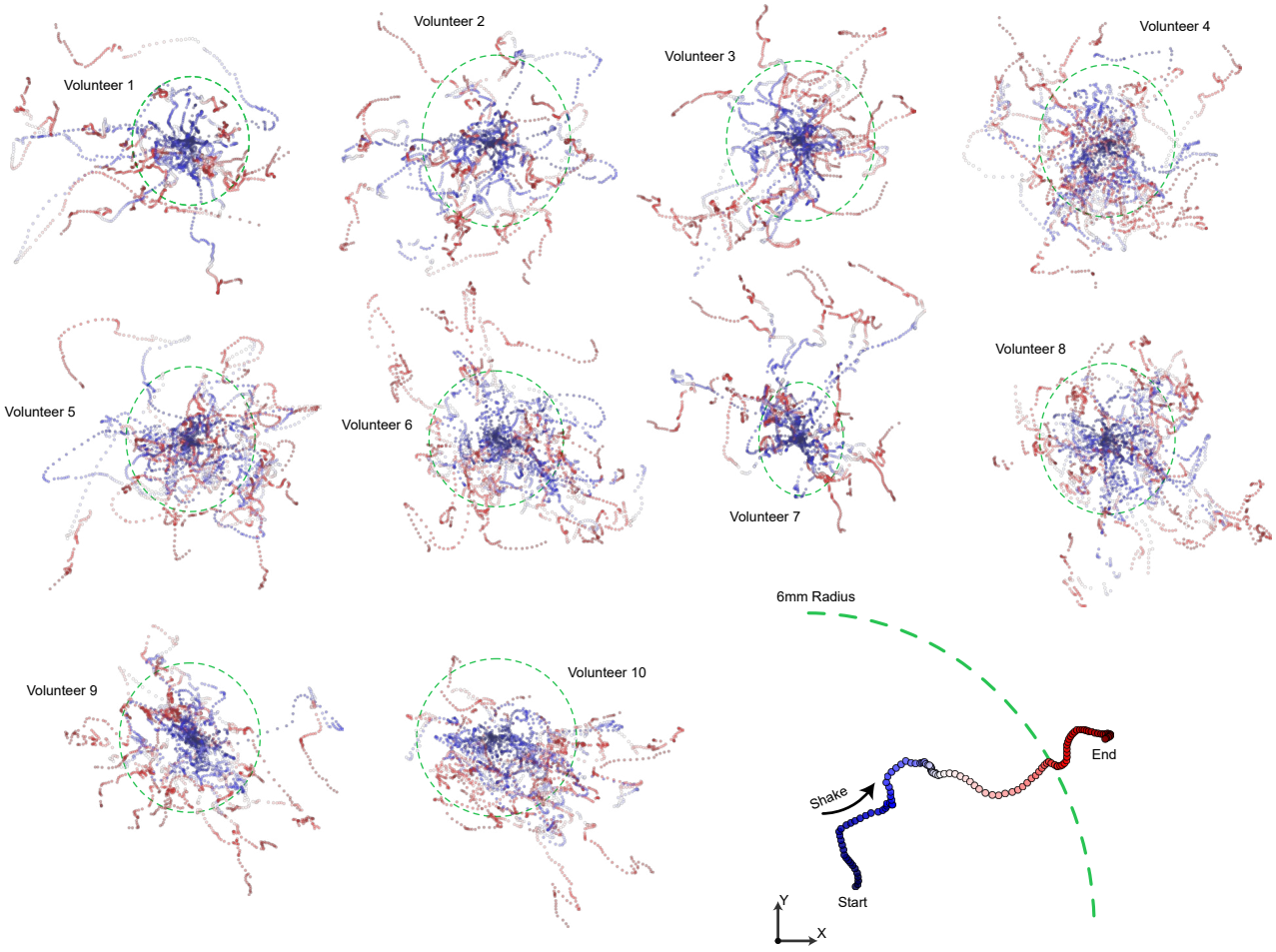

Figure 1: Hand-shake point clouds for individual volunteers. Each hand shake path is recentered to start at the same location, and follows the progression from blue to red over the recorded $N = 120$ frames. Each green dashed oval illustrates a radius of 6mm from the center of the starting point of these hand shake paths, and serves to denote the scale of the point cloud. We find that both the scale and symmetry of hand shake paths is user-dependent, with volunteer 7 exhibiting large asymmetric hand tremors, and volunteer 4 small relatively symmetric hand shakes.

1.2. Hand Shake Statistics

Bundle Length N:	120 Frames	100 Frames	80 Frames	60 Frames	30 Frames	15 Frames
Median Effective Baseline [mm]	5.75	5.34	4.60	4.021	2.62	1.5
Fraction of Effective Baselines $>$ 5mm	0.589	0.517	0.461	0.369	0.183	0.06
Fraction of Effective Baselines $>$ 3mm	0.844	0.794	0.756	0.661	0.433	0.19

Table 1: Quantitative analysis of the maximum displacement, or equivalently the effective baseline, of recorded N frame bundles. We investigate the median effective baseline of these bundles, as well as the fraction of bundles which achieve a greater than 5mm and 3mm effective baseline.

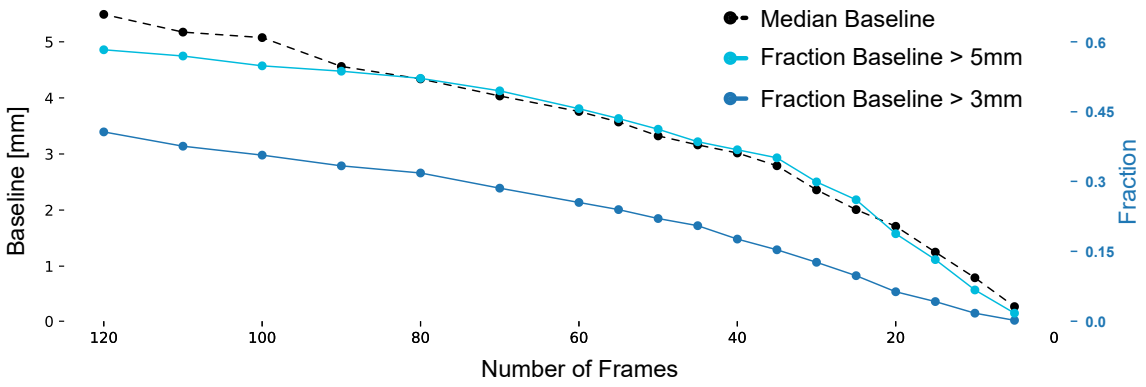

Figure 2: Plot to illustrate the hand shake statistics presented in Table 1. We see that for frame count $120 \geq N \geq 60$ there is not a sharp drop in the fraction of bundles which record a $>$5mm maximum displacement.

Probabilistic View of Hand Shake Bundles. Given the randomness of natural hand tremor present during a snapshot recording, there is a corresponding randomness in the effective baseline of a recorded length N bundle. While we can consider the direct correlation between N and phone displacement during recording, where we expect the phone to move farther as we record longer, we offer a more probabilistic view of this process. Given a target for micro-baseline depth reconstruction, for example an effective 5mm or 3mm baseline, we can ask with what probability we expect a length N recorded bundle to provide meet this target. Looking to Table 1, we find that for our collected hand shake data $N = 100$ frames is sufficient to record a 5mm maximum displacement approximately half the time. As seen in Figure 2, bundles between $N = 60$ and $N = 120$ frames in length offer a reasonable probability ($> \frac{1}{3}$) of capturing $>$5mm baseline data. Thus, much like in multi-frame superresolution [2, 3], we can limit the amount of data we acquire during one snapshot bundle if we expect the phone photographer to take multiple snapshots of an object of interest – looking for the perfect angle, a sharp photo – one of which we expect to achieve our target baseline.

2. Additional Results

2.1. Additional Scenes

Figure 3 shows reconstruction results for five additional captured scenes.

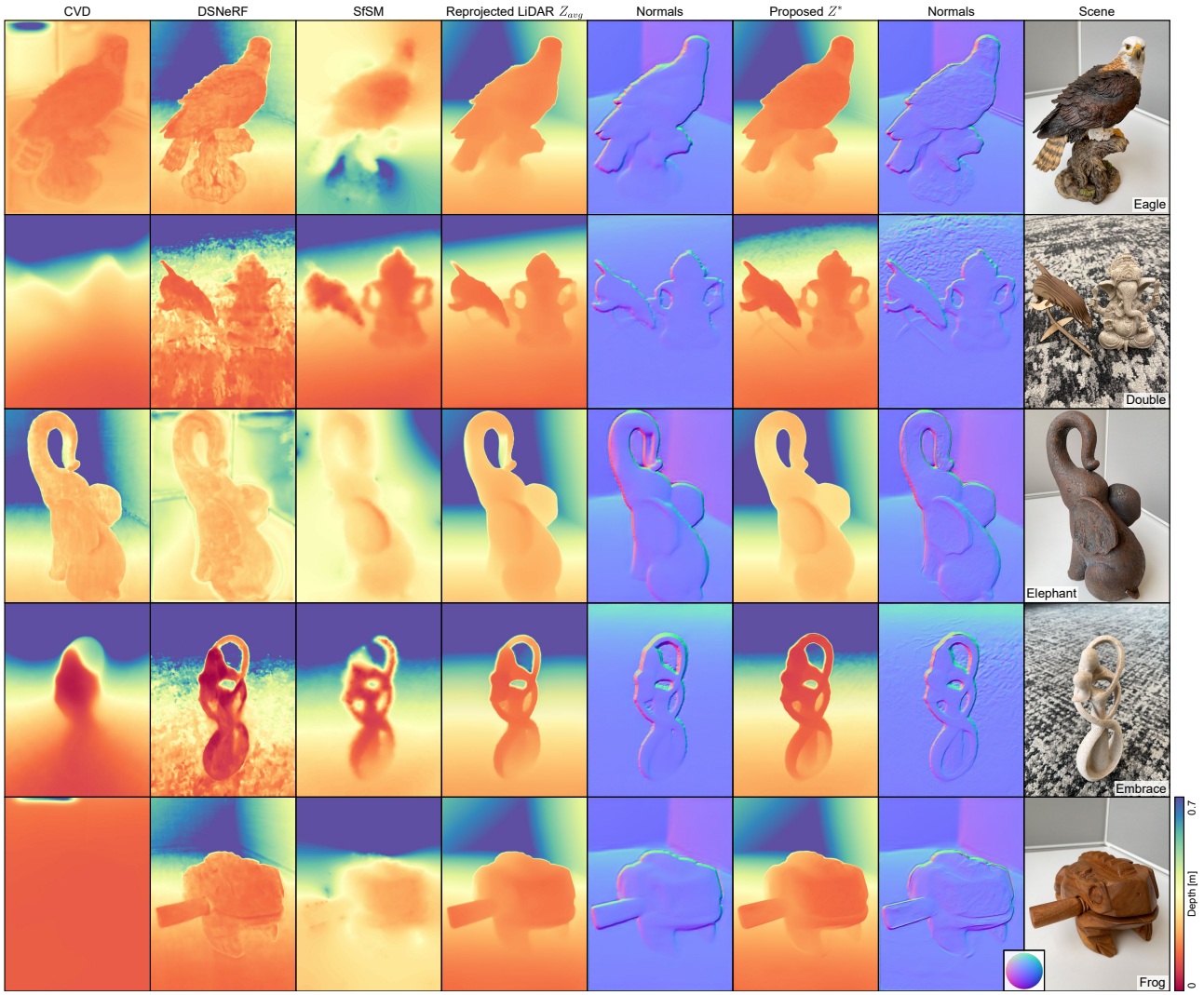

Figure 3: Qualitative comparison of depth reconstruction methods for five additional tabletop scenes (*eagle*, *double*, *elephant*, *elephant*, *frog*). Of note is our recovery of fine depth details in the complex *double* scene, correctly reconstructing the wooden legs of the dolphin statue, and how our proposed method fixes the artifacts in the underling LiDAR depth around the trunk of the *elephant* example. These examples also demonstrate how our proposed method can reconstruct objects at varying scales, where the *embrace* statue is only 10cm tall while the *elephant* is a full half meter in height.

2.2. Photometric Error Maps

Figure 4 visualizes squared photometric error for our proposed method and the next best reconstruction.

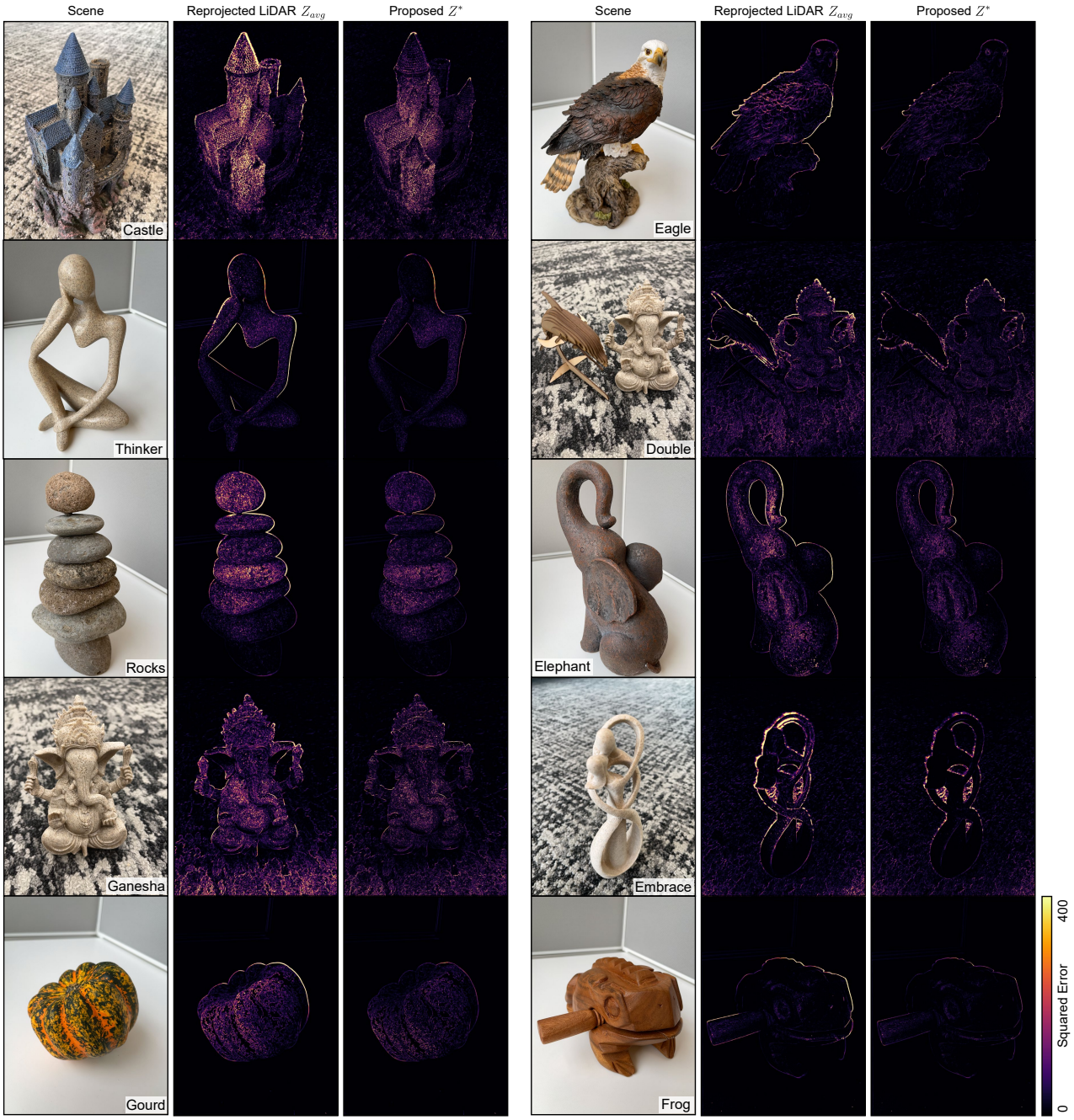

Figure 4: Visualization of squared photometric error PE, as defined in Section 5 of the main text. We see that nearly all the photometric error is accumulated from the high image contrast areas (e.g. object borders and large color features). This matches our previous intuition, as these are the regions our proposed method refines. We see a significant reduction in error around depth discontinuities, such as around the borders of the objects, as well as intra-object features such as the feathers of *eagle*, the central curves of *gourd*, and the body of *ganesha*.

2.3. Long-Range Scenes

Figure 5 shows reconstruction results for captured room-scale scenes outside the main operating range of our method (1+ meters).

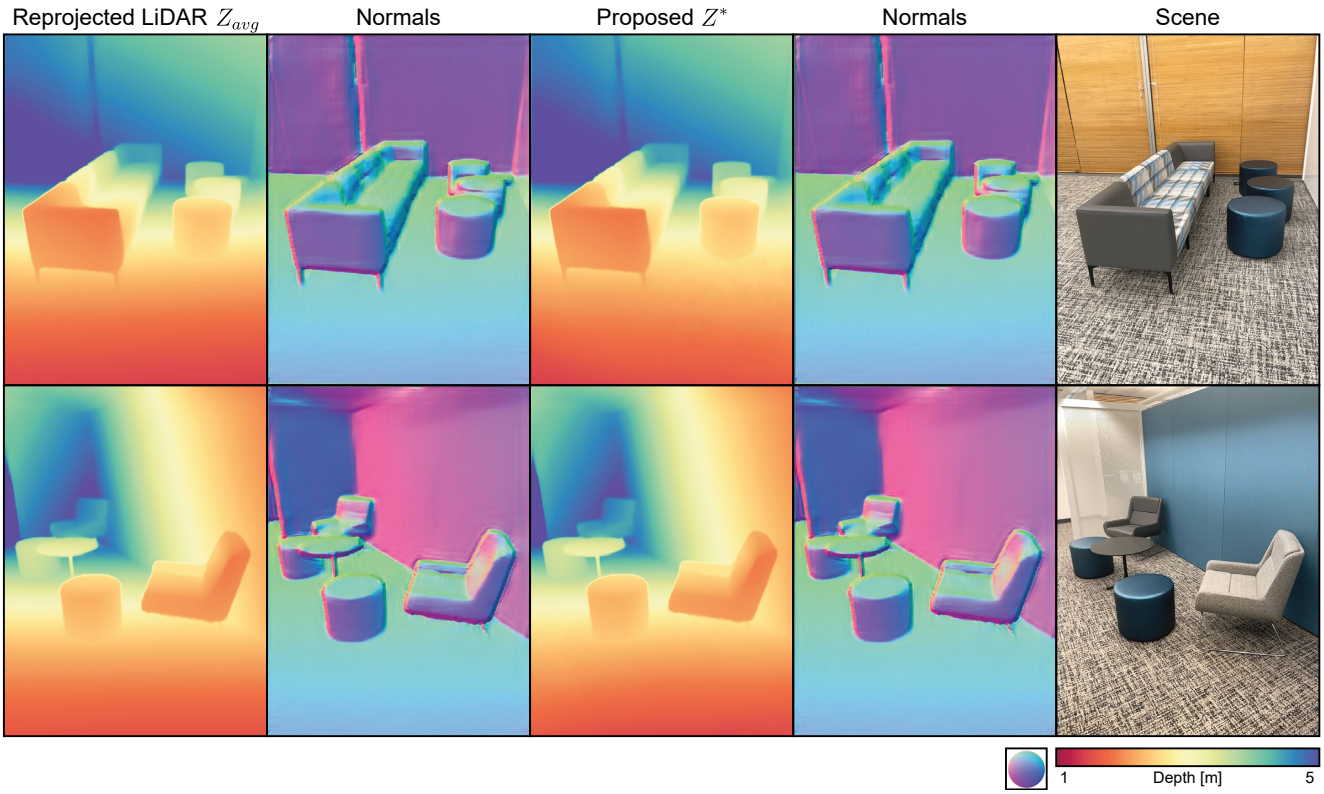

Figure 5: As outlined in Section 4 of the main text, even well-textured objects at depths of greater than 1 meter produce sub-pixel disparities for a micro-baseline stereo setup. In this scenario there is insufficient photometric guidance to reconstruct any new depth features, and our results gracefully degrade to the averaged LiDAR depth output. As we learn depth offsets Δz rather than the direct depth z, our model can trivially output no offset – unperturbed by the minimal photometric loss.

2.4. Close-Range Scenes

Figure 6 contains a set of captured scenes where the majority of the object was out of focus, closer to the phone camera than its minimum focus distance.

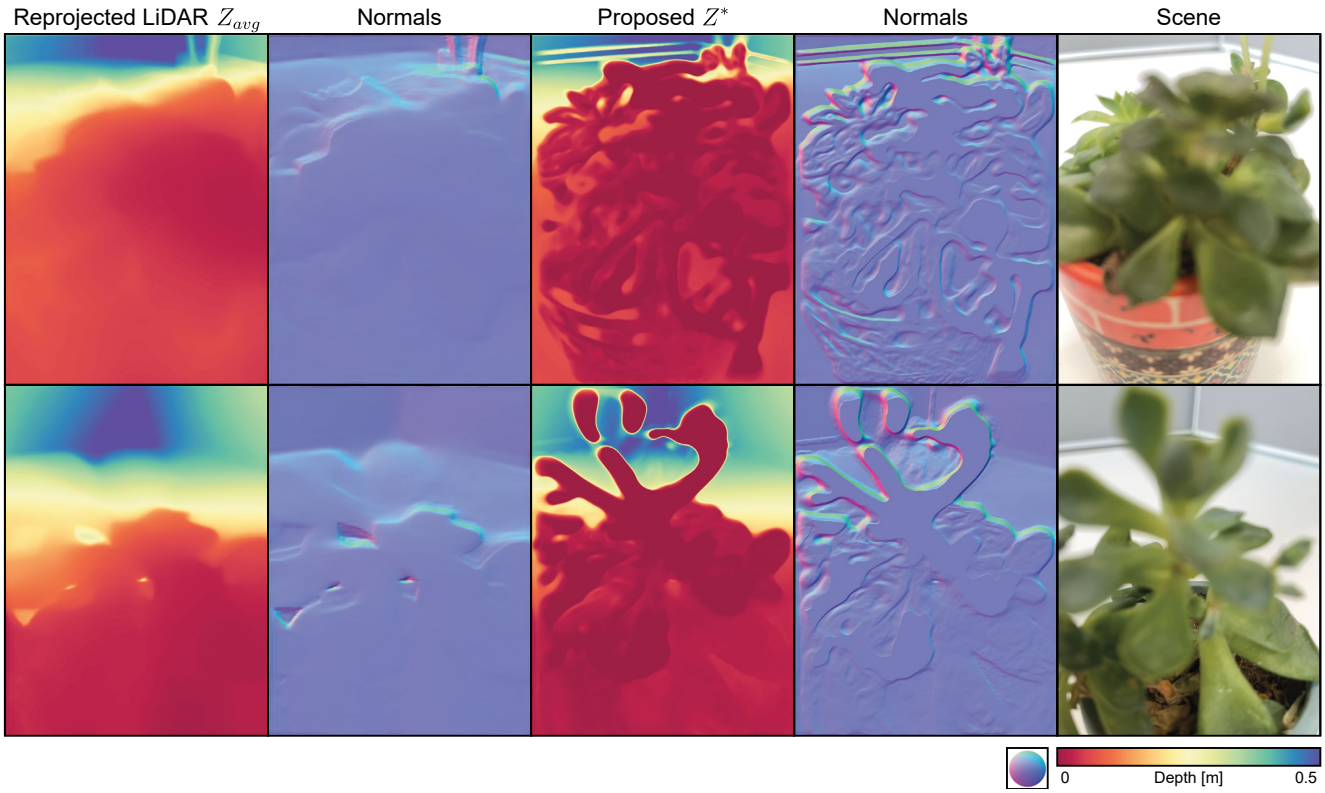

Figure 6: While closer-range objects provide better parallax cues from which to reconstruct depth, our phone has a minimum focus distance of approximately 8cm. Objects closer than this exhibit significant lens blur as can be seen in the scene photos above. In this regime, useful photometric information is overpowered by lens distortions and the now mixed background and foreground pixel colors. While our method can recover depth features for in-focus regions (e.g. the flower pot and lower petals), neither our method nor the underlying LiDAR depth data is able to produce reliable depth for regions under extreme blur. In practical application we expect this to be a rare edge case as a phone photographer is unlikely to capture a snapshot of scene which is clearly blurry in the viewfinder. One potential workaround, implemented by many popular camera applications, is to produce a notification which urges the user to "move back to improve focus".

2.5. Mixed-Range Scenes

Figure 7 demonstrates what happens when a scene contains both close-range (<0.5m) and distant (>1m) content.

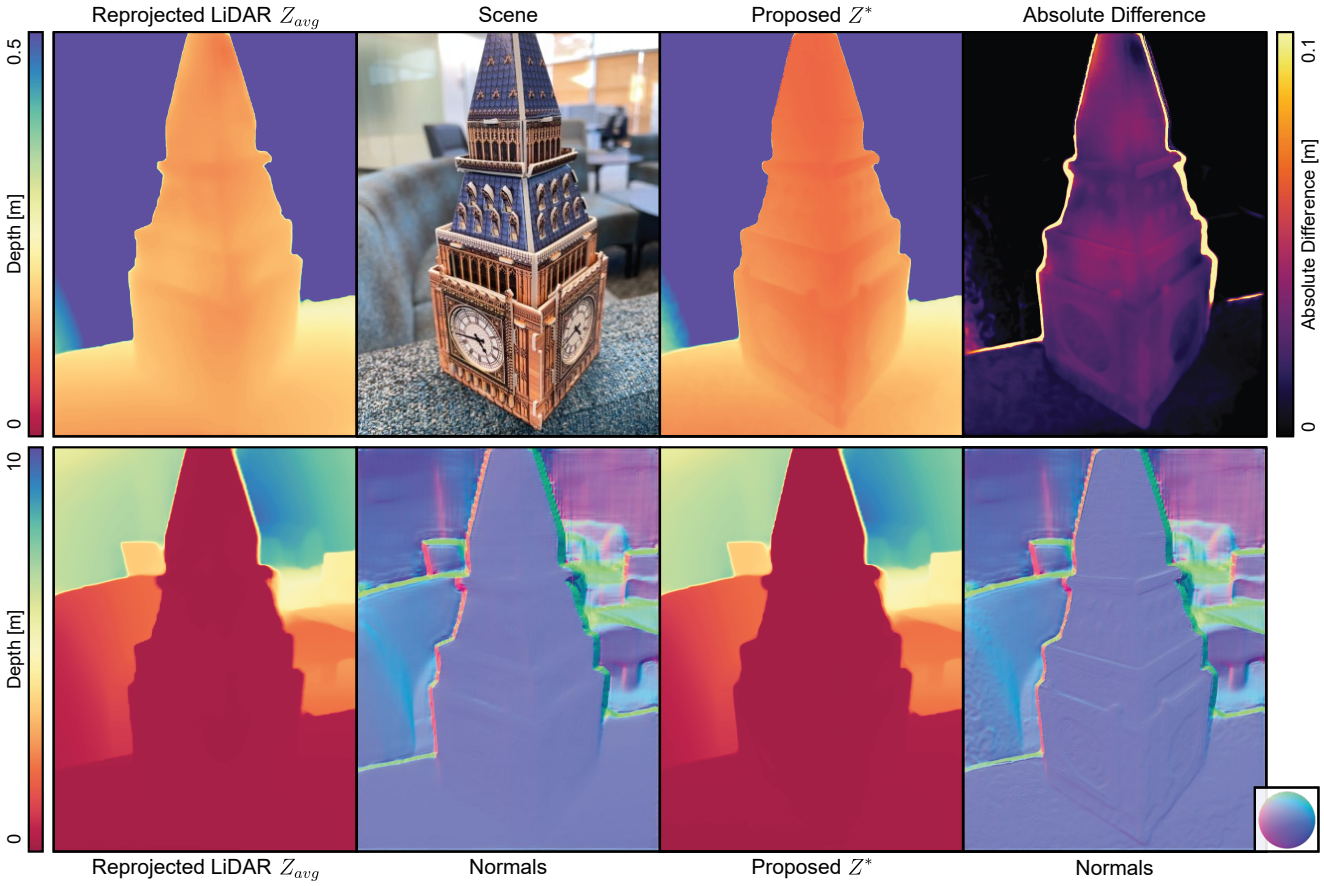

Figure 7: In this scene we have an object resting on a couch arm approximately 15cm away from the phone, with the background content of the room extending 8+ meters into the distance. We note that the short focus distance also noticeably blurs these distant regions. Similar to what is observed in Figure 6, the depth of the background regions remains unperturbed, as it provides virtually no meaningful photometric information. At the same time, the foreground object is successfully reconstructed and we recover fine geometric features such as bumps and protrusions on the surface of the replica clock-tower. This demonstrates that our method, by virtue of its design alone, refines close-range objects without adversely affecting their un-refineable surroundings.

2.6. Reconstructions for Variable Scale of Hand Shake

Figure 8 illustrates how depth reconstruction is affected by the maximum effective baseline of a bundle capture.

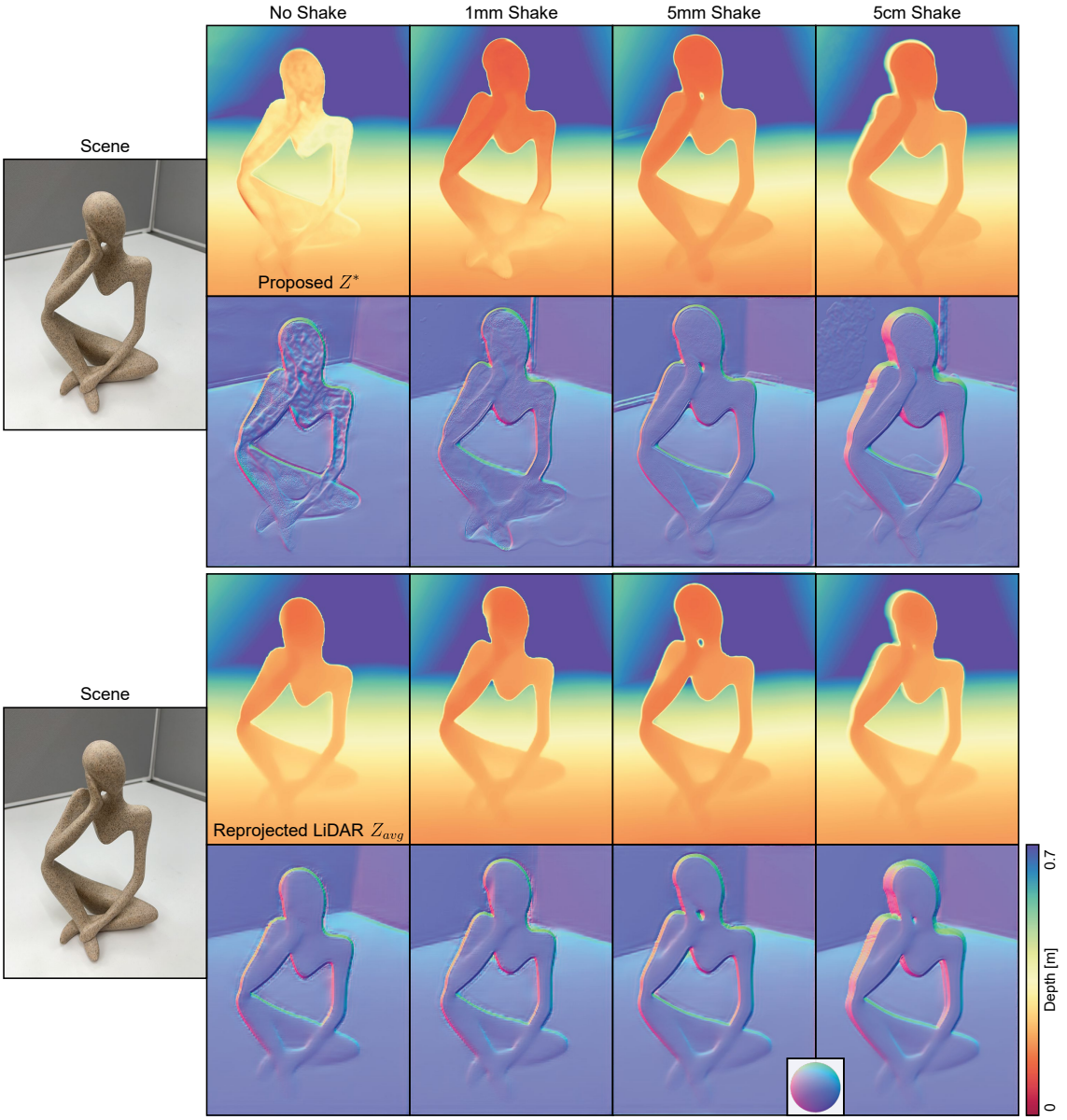

Figure 8: As discussed in Section 1.2 of this supplemental document and Section 4 of the main text, the maximum effective baseline of a recorded bundle is important in determining what we are able and unable to reconstruct. As seen above, with a static camera (and thus an effective baseline of 0mm), there is not only insufficient parallax information from which to reconstruct new depth features, but nearly all the information we do have is incorrect. Small errors in the pose estimation algorithm dominate the overall signal, and we reconstruct incorrect and noisy depths. As we increase the effective baseline to 1mm we see the effects of this noise diminish, however there still appear to be depth artifacts where we find incorrect local minima solutions due to the small view variation. At 5mm we see recovery of fine depth details consistent with Figure 3 and results in the main text, this is what is considered an example of a *good hand shake*. When the effective baseline is further increased to 5cm (closer to a typical stereo camera setup), we see some improvement in reconstruction of fine features such as the model's feet, however both our and the reprojected lidar depth produce a large number of flying pixels around the edges of the statue. This is due to the fact that neither model can explicitly handle occlusions, and instead mix foreground and background depths around discontinuities.

2.7. RGB-Guided Upsampling Results

Figure 9 shows results for an RGB-guided depth upsampling approach Deformable Kernel Networks (DKN) [1] when used on our reconstructed depth data.

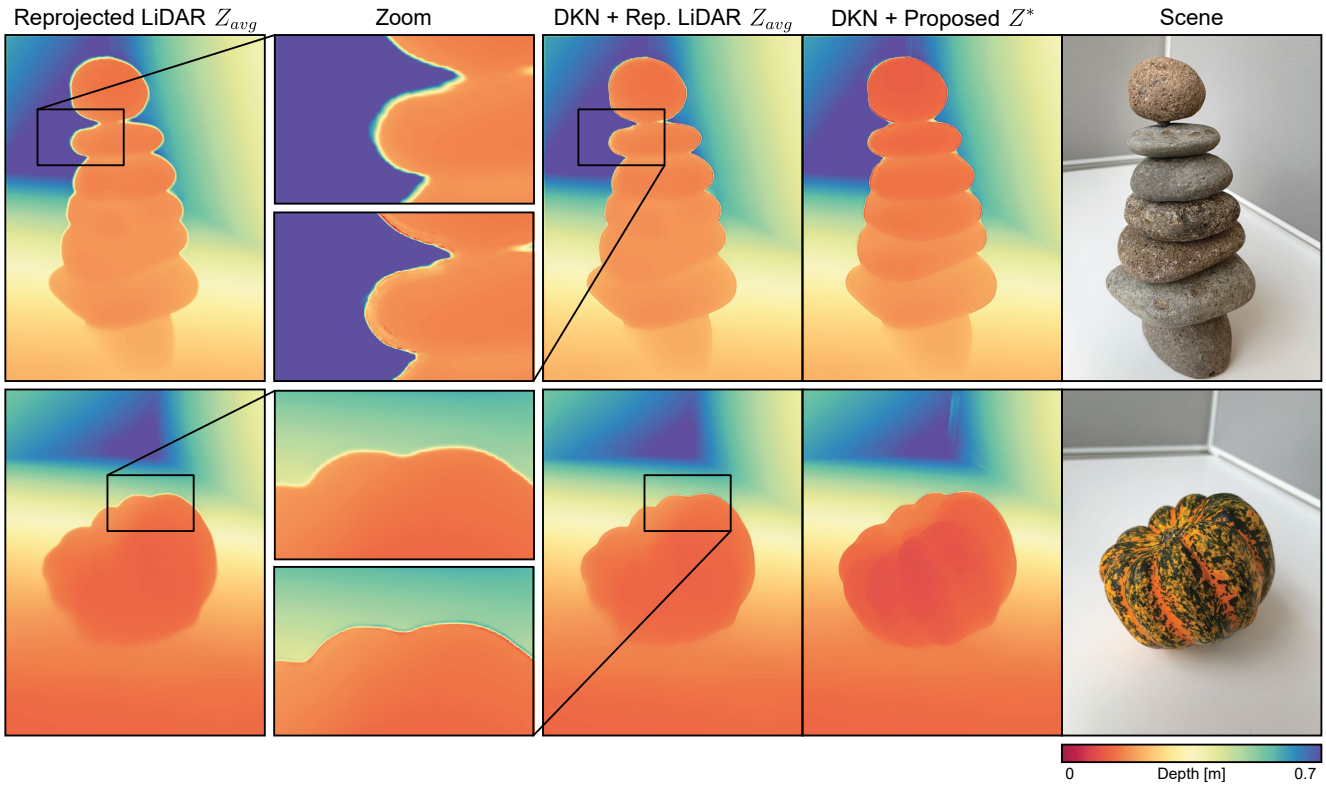

Figure 9: RGB-guided depth upsampling methods use a high-resolution RGB image as guidance for the superresolution and refinement of potentially noisy depth data. These leverage the prior that depth and color changes are often highly correlated to produce visually consistent high-resolution depth outputs. We find that for our data that DKN can use this correlation to improve the reconstruction of sharp depth edges, matching them to the edges of objects in the image, as seen in the zoomed views above. This can often correct for the edge blurring caused by occlusions as discussed in Section 2.6. However, as RGB-guided upsampling does not take parallax effects into account, this approach cannot extract any new depth features that were not already present in the underlying data. Correspondingly, areas not along an edge see very little change in depth after the application of DKN.

3. Ablation Experiments

3.1. Patch Size Ablation

Figure 10 illustrates changes in reconstruction for two representative scenes produced by varying the patch size K used to calculate photometric loss during training, as outlined in Section 3 of the main text.

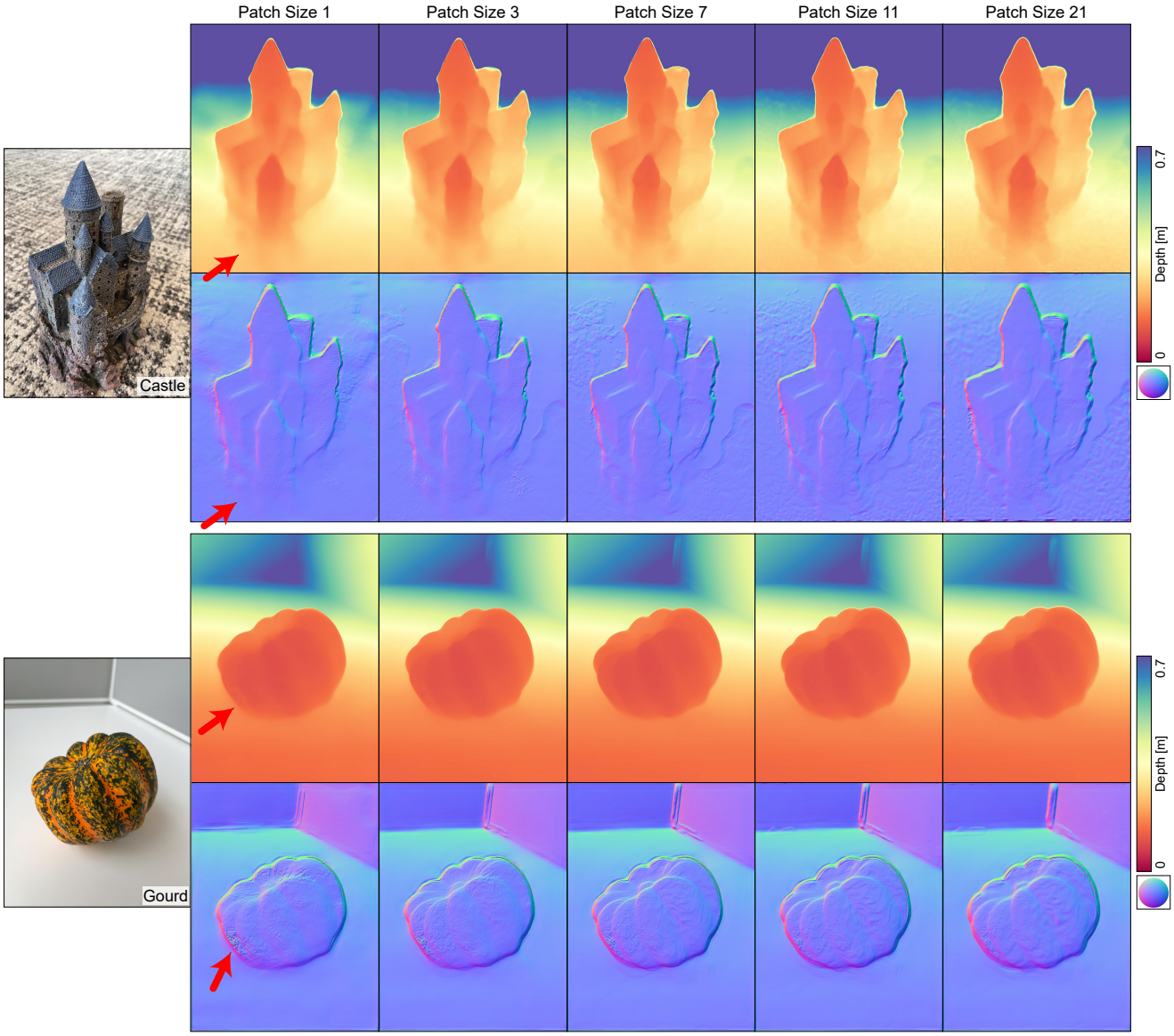

Figure 10: Reconstruction results for varying patch size K. Note the depth artifacts, highlighted with red arrows, where single-pixel photometric information is insufficient to correctly resolve depth, and false matches drive the reconstruction to find incorrect local minima solutions. As we increase patch size these artifacts disappear, but we also fatten the reconstructed object edges and can blur thin structures. Patch size 11 achieves a reasonable mezzanine between artifact suppression and fine feature reconstruction.

3.2. Regularization Weight Ablation

Figure 11 shows reconstruction results for our method with varying geometric regularization weight α.

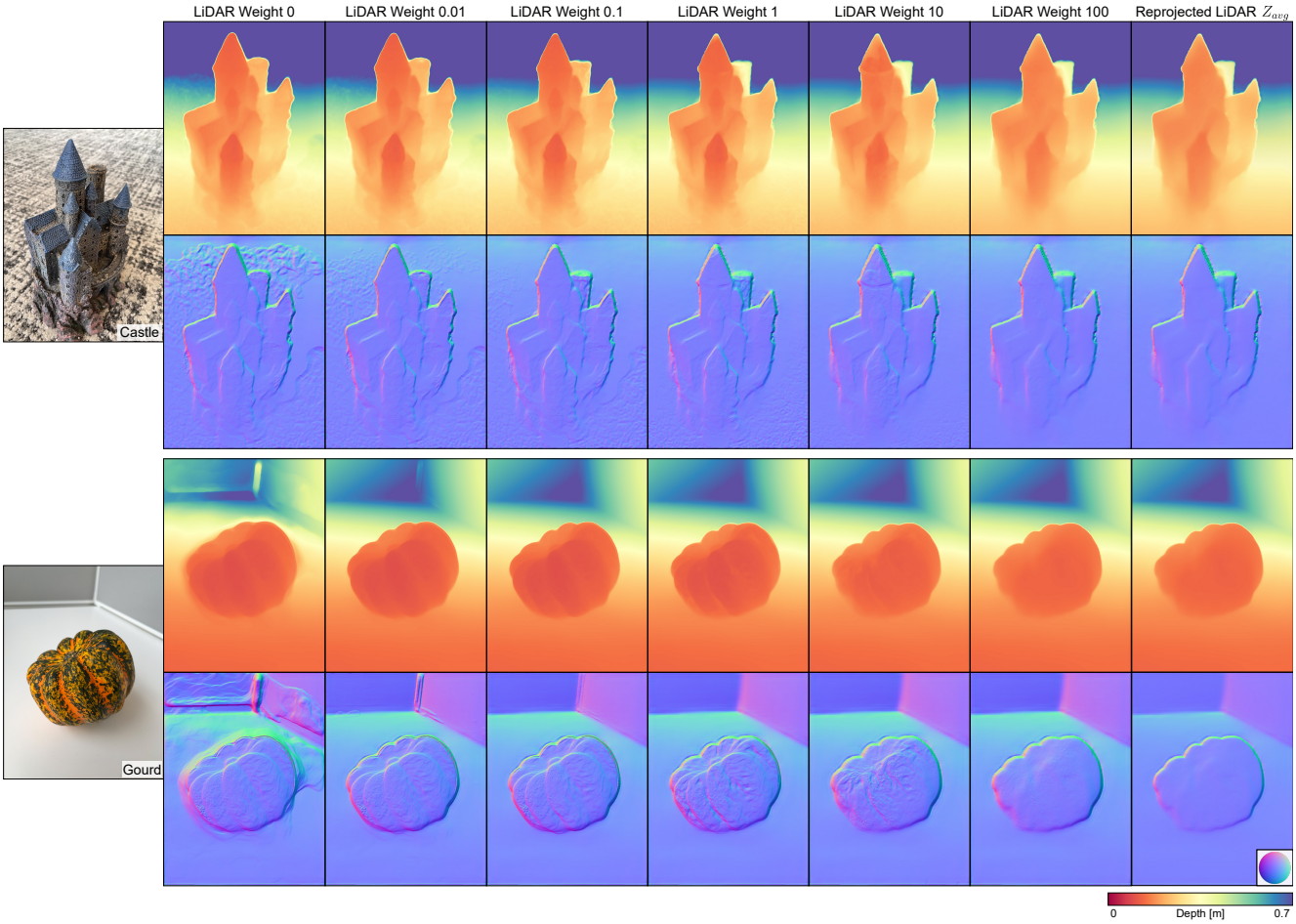

Figure 11: Reconstruction results for varying geometric regularization weight α. With no geometric regularization term ($\alpha = 0$) we see that the depth reconstruction of the textureless object backgrounds erroneously drift away from the underlying LiDAR data as there is no guiding photometric loss in those regions. When α is non-zero, these textureless regions are effectively locked to the LiDAR depth solution. As we increase α, the overall reconstruction is pulled closer to the LiDAR depth, with the reconstructed depth for $\alpha = 100$ being virtually identical to the reprojected LiDAR depth Z_{avg}. As we desire a reconstruction almost entirely guided by micro-baseline photometric information, to fix LiDAR depth artifacts like those at the top of the *castle* example, we choose a small $\alpha = 0.01$.

3.3. Encoding Function Ablation

Figure 12 illustrates changes in reconstruction for our method produced by varying positional encoding function count L.

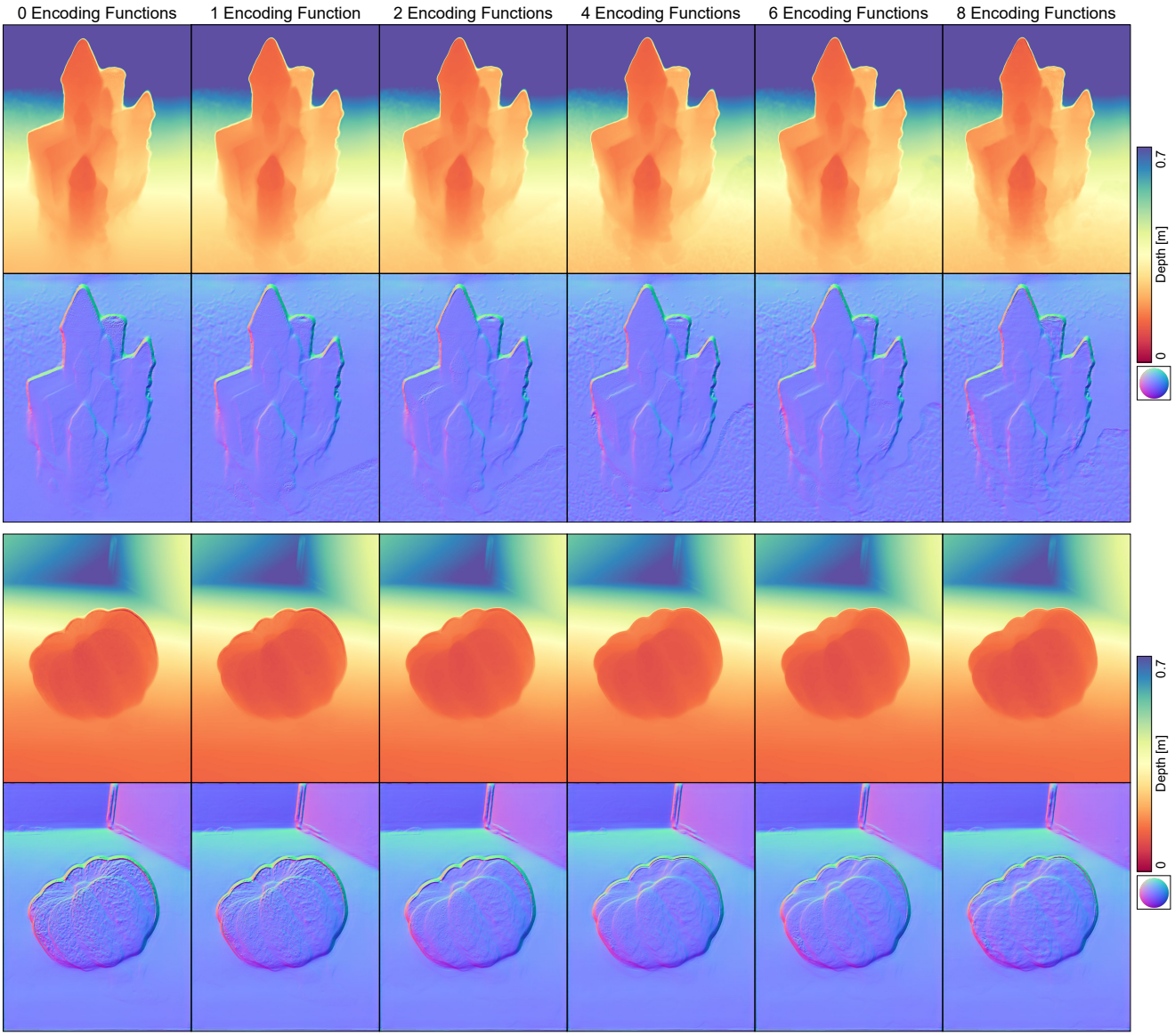

Figure 12: Reconstruction results for varying encoding function count L. We see that without positional encoding, for $L = 0$, our reconstruction is quite noisy and does not correct for the depth artifact at the top of the *castle* example. We speculate the MLP overfits to RGB and depth information in this case, rather than learning true corrective offsets for depth refinement. At $L = 8$ encoding function we start to see high frequency artifacts bleed into the reconstruction. We thus choose $L = 6$ functions as a reasonable mezzanine where neither of these effects are present.

3.4. Gaussian Versus Square Patch Ablation

Figure 13 highlights the effects of Gaussian weighted patches, as described in Section 3 of the main text, as compared to uniformly weighted *square* patches.

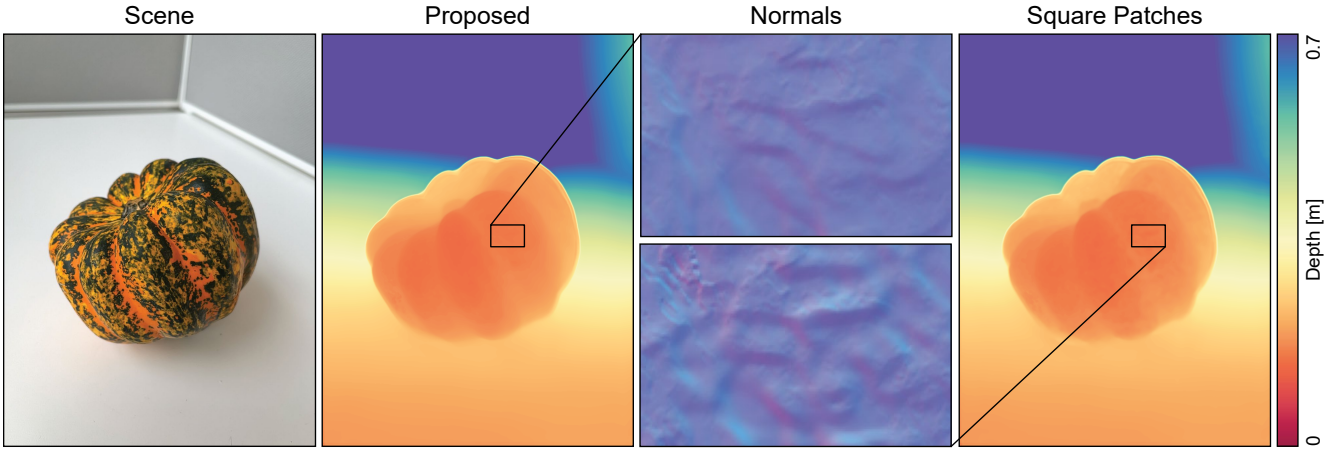

Figure 13: In the micro-baseline problem setup we have very little view variation between frames in a captured bundle. Thus patch-matching, as shown in Section 3.1, is important to avoid returning incorrect depth solutions for uniformly colored patches of the image space. Large square patches, however, uniformly integrate signal from potentially non-planar surfaces and introduce noise when the actual deformation of these pixels doesn't match our planar-warp assumption. The Gaussian kernel weighting emphasizes correctly matching the center of these image patches (where this warping would be minimal), but still provides the edge information for situations where the center might not contain salient color information for matching. This results in overall smoother reconstructions, as can be seen in the normal maps above.

References

[1] B. Kim, J. Ponce, and B. Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021. 10

[2] R. Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 3

[3] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 3