

# SH-SAS: An Implicit Neural Representation for Complex Spherical-Harmonic Scattering Fields for 3D Synthetic Aperture Sonar

## Supplementary Material

The supplemental materials include several ablations and additional results that extend the findings of the main manuscript. In particular, we provide additional simulated reconstructions with qualitative comparisons and per-object visualizations. An SH-level ablation study explores the effect of varying the maximum degree  $L \in 1, 2, 3$ . We also include a threshold sensitivity analysis for iso-surface extraction with marching cubes, highlighting reconstruction stability under different thresholds. In addition, we present signal fitting and novel-view synthesis results, including time-of-flight fitting curves. In addition, code, datasets and visualizations are available at the project website: <https://omkarv23.github.io/SH-SAS-website/>.

### 9. Ellipsoidal Sampling: Derivation and Ray-Ellipsoid Intersection

Our measurements are bistatic time-of-flight (ToF), so each sample constrains contributing scene points to a two-focus ellipsoid with foci at the TX/RX. The correct way to associate a ToF bin with spatial hypotheses is therefore *ellipsoidal sampling*: for a given bin, we sample points where a TX ray intersects its constant-ToF ellipsoid. The derivation below (i) parameterizes that ellipsoid from the bin center, (ii) introduces a canonical “ellipsoid frame” that accommodates arbitrary TX/RX poses, and (iii) provides a closed-form ray–ellipsoid intersection used throughout our pipeline.

**Constant-ToF ellipsoid.** Let  $\mathbf{o}_T, \mathbf{o}_R \in \mathbb{R}^3$  be TX/RX positions,  $d = \|\mathbf{o}_T - \mathbf{o}_R\|$  their separation,  $c$  the sound speed, and  $t$  the ToF of a sample. The locus of points with

$$t = \frac{R_T + R_R}{c}, \quad R_T = \|\mathbf{x} - \mathbf{o}_T\|, \quad R_R = \|\mathbf{x} - \mathbf{o}_R\|$$

is an ellipsoid with foci at  $\mathbf{o}_T, \mathbf{o}_R$  and semi-major

$$a = \frac{ct}{2}, \quad b = c = \sqrt{a^2 - (d/2)^2}.$$

**Ellipsoid frame.** Let  $\mathbf{m} = \frac{1}{2}(\mathbf{o}_T + \mathbf{o}_R)$  and choose  $R \in SO(3)$  so that  $R^\top(\mathbf{o}_R - \mathbf{o}_T) = (d, 0, 0)^\top$ . Define ellipsoid-frame coordinates

$$\tilde{\mathbf{x}} = R^\top(\mathbf{x} - \mathbf{m}), \quad \tilde{\mathbf{o}}_T = R^\top(\mathbf{o}_T - \mathbf{m}), \quad \tilde{\mathbf{d}}_T = R^\top \mathbf{d}_T,$$

where  $\mathbf{d}_T$  is a unit TX ray direction. In this frame, the ellipsoid is axis-aligned:

$$\frac{\tilde{x}^2}{a^2} + \frac{\tilde{y}^2}{b^2} + \frac{\tilde{z}^2}{b^2} = 1.$$

**Ray–ellipsoid intersection.** A TX-emitted ray is

$$\tilde{\mathbf{x}}(l) = \tilde{\mathbf{o}}_T + l \tilde{\mathbf{d}}_T, \quad l \geq 0.$$

Plugging into the quadric gives  $a_0 l^2 + b_0 l + c_0 = 0$  with

$$\begin{aligned} a_0 &= \frac{\tilde{d}_{T,x}^2}{a^2} + \frac{\tilde{d}_{T,y}^2}{b^2} + \frac{\tilde{d}_{T,z}^2}{b^2}, \\ b_0 &= 2 \left( \frac{\tilde{o}_{T,x} \tilde{d}_{T,x}}{a^2} + \frac{\tilde{o}_{T,y} \tilde{d}_{T,y}}{b^2} + \frac{\tilde{o}_{T,z} \tilde{d}_{T,z}}{b^2} \right), \\ c_0 &= \frac{\tilde{o}_{T,x}^2}{a^2} + \frac{\tilde{o}_{T,y}^2}{b^2} + \frac{\tilde{o}_{T,z}^2}{b^2} - 1. \end{aligned}$$

If  $\Delta = b_0^2 - 4a_0c_0 < 0$ , there is no intersection for that ToF. Otherwise,

$$l_{\pm} = \frac{-b_0 \pm \sqrt{\Delta}}{2a_0}, \quad l^* = \min\{l_{\pm} \mid l_{\pm} > 0\},$$

and the sample location is  $\mathbf{x}^* = \mathbf{m} + R \tilde{\mathbf{x}}(l^*)$ . (Use the smallest positive root since the ray originates at the TX.)

### 10. Additional Experimental Results

**Additional Results: Simulated Data.** Qualitatively, our reconstructions are cleaner and more complete across all four synthetic scenes (Armadillo, Bunny, Happy Buddha, XYZ Dragon) Fig. 6: backprojection exhibits speckle and fragmentation, while Reed et al. tends to over-smooth and lose thin structures; our method recovers sharper geometry with fewer holes and floaters. Quantitatively, our approach ranks best on *most* metrics and is never worst. We evaluate the methods on the Chamfer distance, IoU, precision, and F1 score. On average, our method outperforms the baseline methods as shown in Tab. 3.

On simulated data we also reach target quality in fewer iterations, yielding shorter overall training; Fig. 7 illustrates that our reconstructions achieve higher quality earlier in training.

#### Additional Results: AirSAS Armadillo and Bunny Data.

We evaluate on real AirSAS measurements for Armadillo and the Stanford bunny (Fig. 8 and 9). In this setting, reconstruction is more challenging due to measurement noise and speckle, residual calibration errors, and stronger shadowing from the single-look, limited-aperture acquisition. Backprojection produces fragmented surfaces and significant floor or pedestal artifacts, and Reed et al. [43] reduces some of this clutter but still exhibits bleed-through

and distorted thickness in the recovered shapes. In contrast, our method yields more compact, coherent meshes that better match the expected silhouettes across views, while noticeably suppressing spurious structure beneath the targets. These results are consistent with the trends observed in simulation, suggesting that the learned SH-based scattering representation transfers to real measurements under the same acquisition geometry.

## 11. Ablation Study: SH levels

**Background on spherical harmonics.** Spherical harmonics (SH) form an orthonormal basis on the unit sphere  $\mathbb{S}^2$  and let us represent directional or view-dependent terms (e.g., scattering/BRDF lobes) as a band-limited expansion. An SH level  $L$  includes all degrees  $\ell = 0, \dots, L$ , for a total of  $(L + 1)^2$  coefficients per scalar field; increasing  $L$  increases the angular bandwidth that can be expressed. Intuitively,  $L=1$  captures a diffuse term plus first-order lobes,  $L=2$  adds quadratic variation, and  $L=3$  can express noticeably sharper, asymmetric lobes. In our pipeline, these coefficients are learned at each 3D location and evaluated along ellipsoidal samples consistent with the bistatic ToF geometry.

**Ablation results.** We ablate the SH order by increasing the level from 1 to 3 while keeping all other settings fixed. Fig. 10 show a clear trend: as SH level increases, reconstruction quality improves, with level 3 performing best in our setup.

## 12. Effect of Threshold on Visualization

Figure 11 examines how sensitive each method is to the marching-cubes threshold used to extract meshes from the learned field. As the threshold increases (top to bottom), Backprojection swings from an overfilled slab-like volume to severe erosion and fragmentation, indicating a poorly calibrated field whose level set is not tied to geometry. Reed et al. [43] is more stable but still exhibits large changes: pedestal/floor artifacts at low thresholds. In contrast, Ours is markedly consistent across the changes in threshold and surfaces change mostly by a uniform thinning/thickening, with minimal new holes or floaters. This robustness suggests that our SH-based directional modeling learns a well-calibrated implicit field, making mesh visualization far less sensitive to the particular threshold choice.

## 13. Signal fitting and novel-view transient synthesis

We evaluate the network’s ability to fit measured signals and then use that fit to synthesize transients at unseen poses. We define *novel-view transient synthesis* as generating the complex transient (analytic signal—real and imaginary parts)

at a transmitter/receiver pose within the synthetic aperture where no physical measurement was taken; given a desired TX/RX location, our learned scene representation predicts the time-resolved response that would have been recorded at that pose. This capability enables simulation for sensor/trajectory design and target modeling, *aperture densification* by interpolating missing pings to improve coherent processing, and data augmentation for downstream detection/classification. Related work in transient imaging motivates novel-view synthesis for virtual sensing and augmentation—TransientNeRF [30] primarily operates in the LiDAR image domain, while TransientAngelo [29] targets single-photon LiDAR and predicts normalized photon-count transients—but our focus is on *analytic* (complex-valued) transients at arbitrary bistatic poses, consistent with the ellipsoidal ToF geometry. Operationally, for each target pose ( $\mathbf{o}_T, \mathbf{o}_R$ ) and orientation we form the constant-ToF ellipsoid per bin, compute ray–ellipsoid intersections to obtain geometry-aware samples, evaluate the learned SH-based directional scattering at those samples, and integrate through the forward model to yield the complex transient—identical to training-time analysis-by-synthesis, but evaluated at new poses. On the *Armadillo* dataset, Fig. 12 shows that our predictions better preserve delay structure and phase, relative to [43]; quantitatively, Table 4 reports consistent improvements across apertures for both the real and imaginary components—lower  $\ell_1$  error and lower mean squared error (MSE).



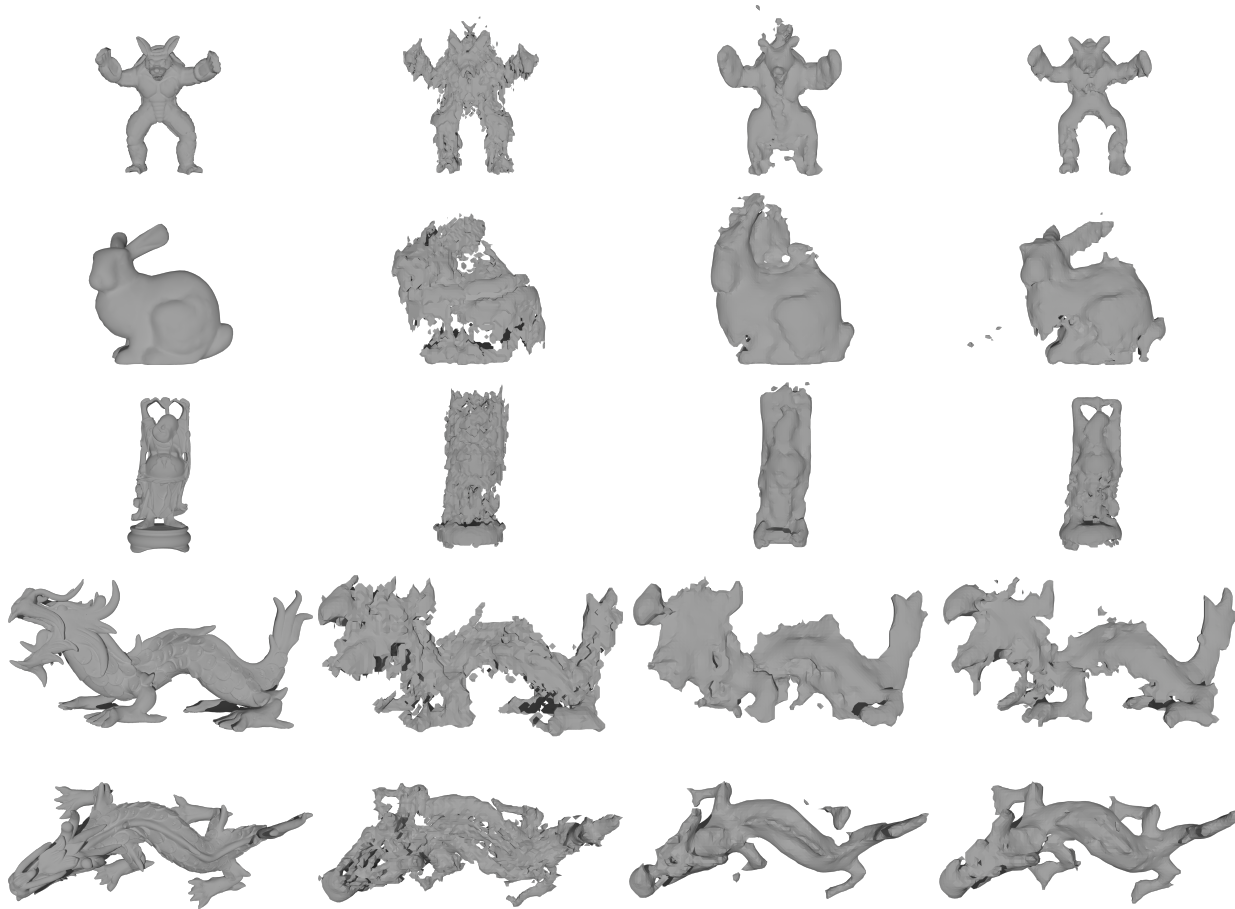


Figure 6. Mesh reconstructions of synthetic data: Armadillo, Bunny, Happy Buddha, XYZ Dragon (side, top) scenes using (from left to right): Ground Truth, Backprojection, Reed et. al [43], and Ours.

Object	Method	Point cloud				Mesh			
		Chamfer	IoU	Precision	F1	Chamfer	IoU	Precision	F1
Armadillo	Backprojection	8.204e-5	0.243	0.280	0.391	1.081e-4	0.223	0.284	0.364
	Reed et al.	7.871e-5	0.418	0.402	0.544	1.209e-4	0.154	0.237	0.266
	Ours	<b>7.009e-5</b>	<b>0.444</b>	<b>0.525</b>	<b>0.614</b>	<b>7.688e-5</b>	<b>0.249</b>	<b>0.393</b>	<b>0.398</b>
Buddha	Backprojection	9.546e-4	0.131	0.132	0.232	1.191e-3	0.056	0.074	0.107
	Reed et al.	4.832e-4	0.601	0.439	0.608	1.184e-3	0.056	0.101	0.106
	Ours	<b>9.238e-5</b>	<b>0.611</b>	<b>0.736</b>	<b>0.756</b>	<b>7.211e-5</b>	<b>0.324</b>	<b>0.550</b>	<b>0.490</b>
Bunny	Backprojection	<b>1.233e-4</b>	0.227	0.325	0.371	<b>1.551e-4</b>	<b>0.197</b>	0.329	<b>0.330</b>
	Reed et al.	1.524e-4	<b>0.412</b>	0.480	<b>0.584</b>	2.190e-4	0.149	0.101	0.260
	Ours	1.255e-4	0.398	<b>0.518</b>	0.569	1.798e-4	0.191	<b>0.368</b>	0.321
Dragon	Backprojection	<b>5.679e-5</b>	0.202	0.238	0.336	8.586e-5	0.167	0.212	0.286
	Reed et al.	6.834e-5	<b>0.360</b>	0.390	<b>0.529</b>	6.306e-5	0.174	0.246	0.296
	Ours	6.752e-5	0.357	<b>0.402</b>	0.526	<b>5.925e-5</b>	<b>0.194</b>	<b>0.270</b>	<b>0.325</b>

Table 3. Core 3D metrics (point cloud vs. mesh) across objects. Chamfer uses scientific notation (3-dec mantissa); others rounded to 3 decimals. Colors indicate rank within each object/column (best=green, second=yellow, worst=red). Ties share the same color.

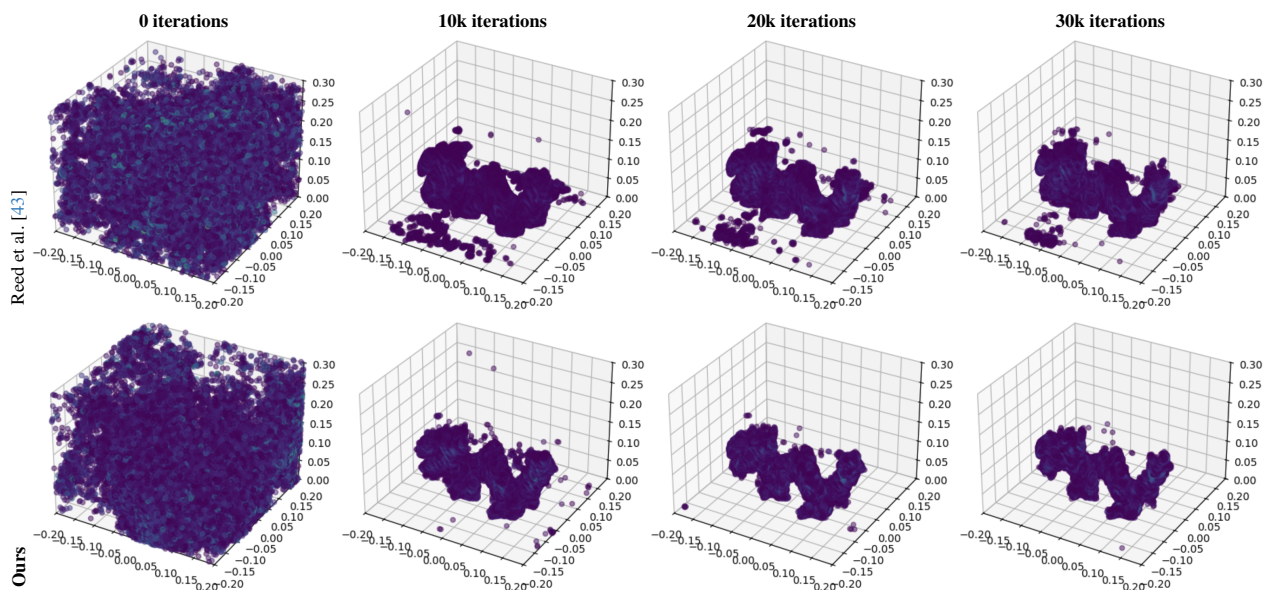


Figure 7. Convergence comparison on XYZ Dragon. Our method (bottom) reconstructs fine-scale structure more quickly and achieves higher quality at earlier iterations compared to Reed et al. [43] (top).

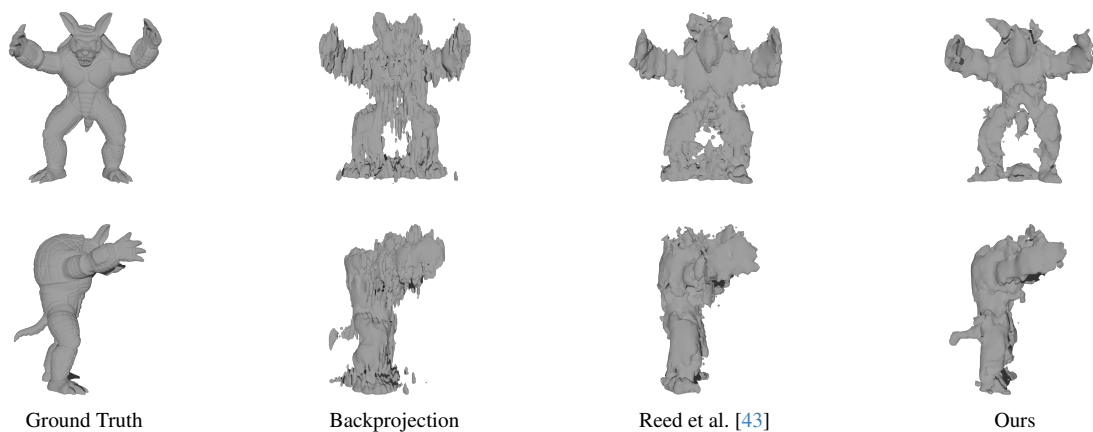


Figure 8. Real data: Armadillo. Top: frontal view. Bottom: side view.

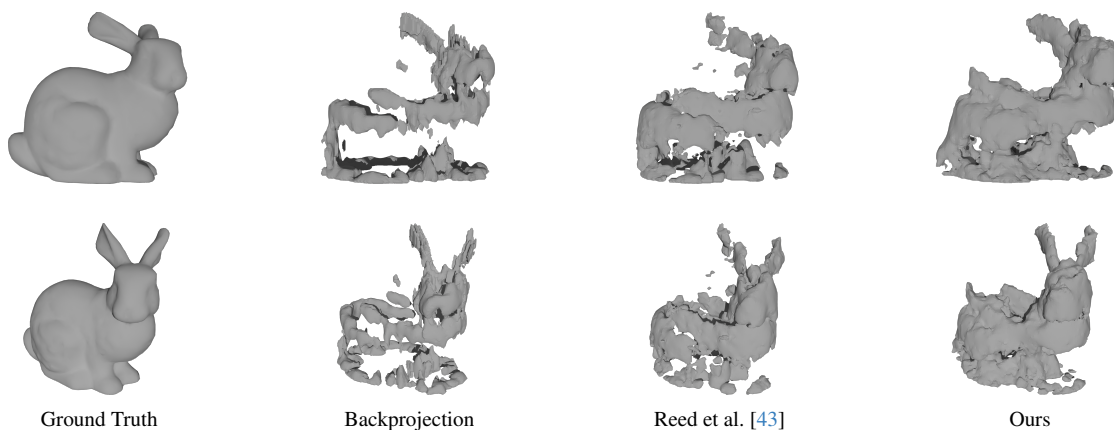


Figure 9. Real data: Stanford bunny. Top: frontal view. Bottom: top view.

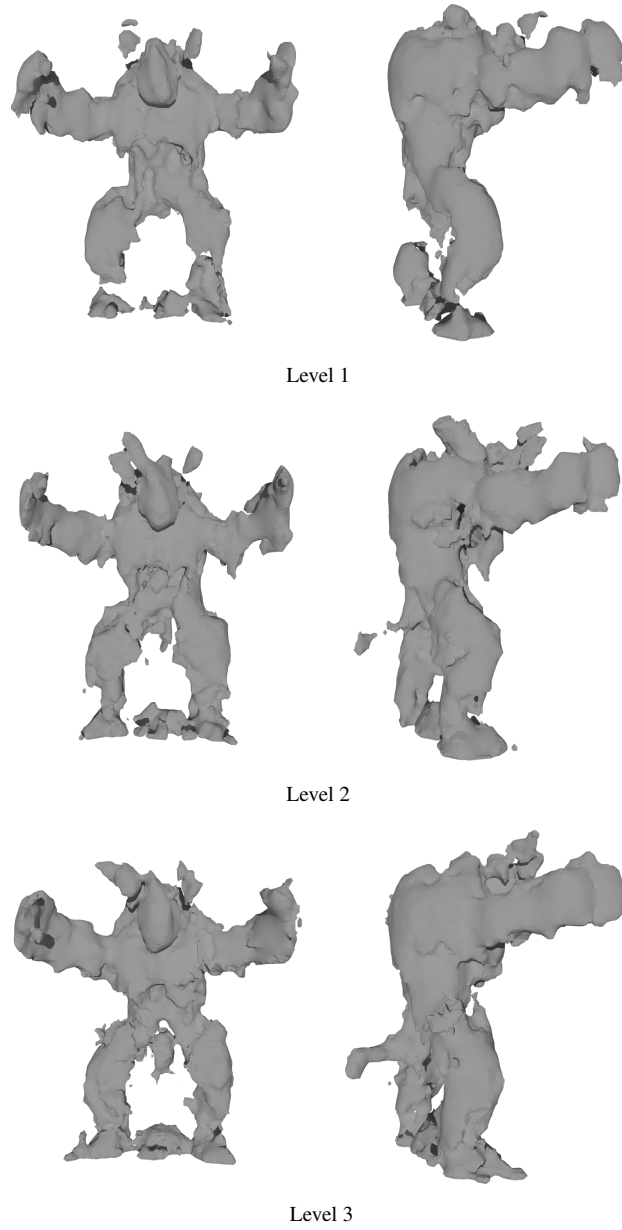


Figure 10. Increase in Spherical Harmonic Levels. As the Spherical Harmonic (SH) level increases from Level 1 to Level 3, the quality of reconstruction improves significantly. Fine details, such as the legs and tail of the object, become clearer and more accurately represented at higher SH levels, demonstrating enhanced fidelity and geometric detail in the model.

Table 4. Novel-view signal errors and GPU render time (lower is better).

Method	Render (ms)	L1 (real)	L1 (imag)	L1 (abs)	MSE (real)	MSE (imag)	MSE (abs)
Reed et al.	147.40	0.417	0.639	0.865	1.339	1.346	2.476
Ours	<b>126.11</b>	<b>0.402</b>	<b>0.576</b>	<b>0.665</b>	<b>0.797</b>	<b>0.810</b>	<b>1.171</b>

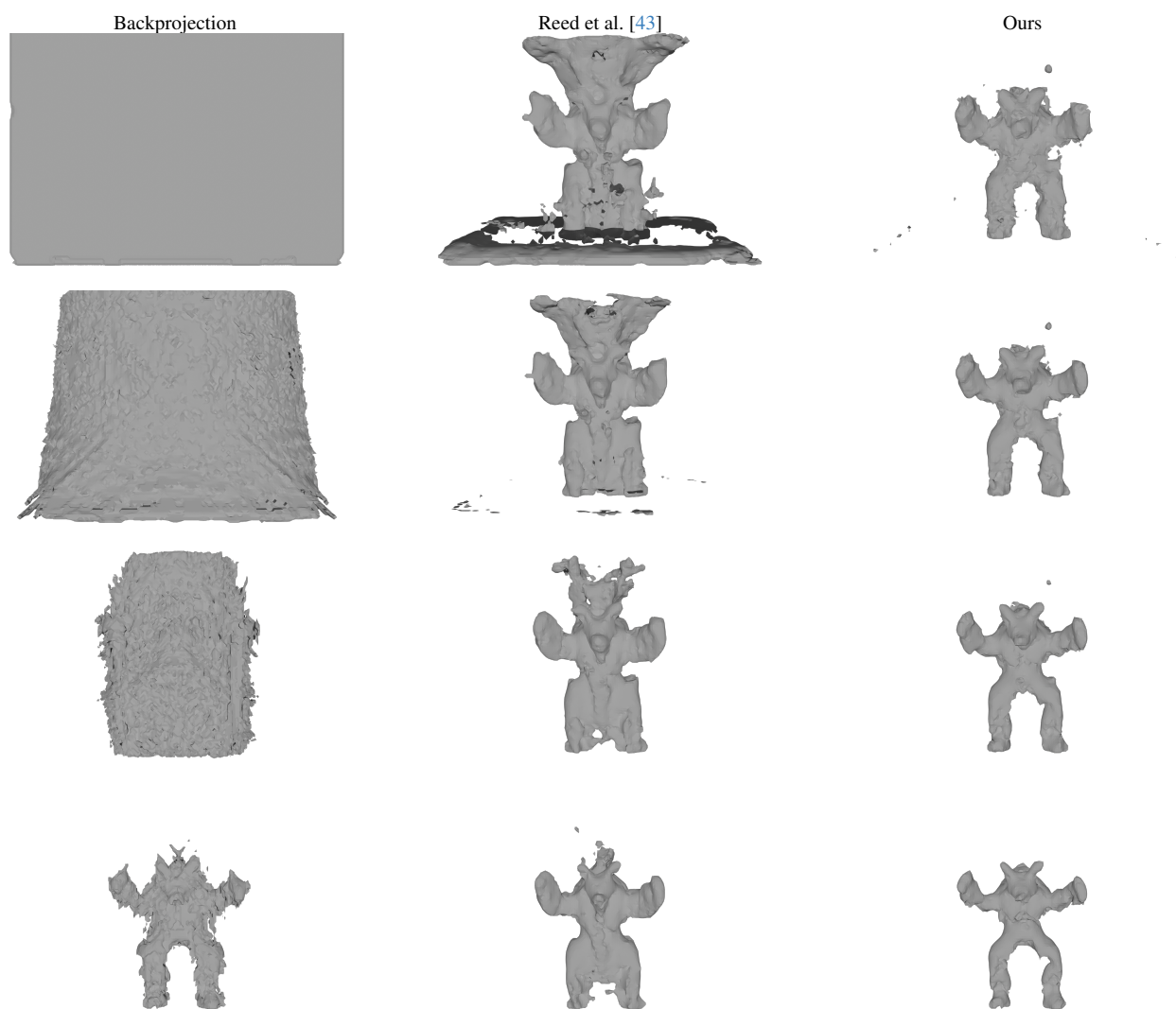


Figure 11. Mesh consistency across threshold levels. Columns: methods. Rows: increasing threshold

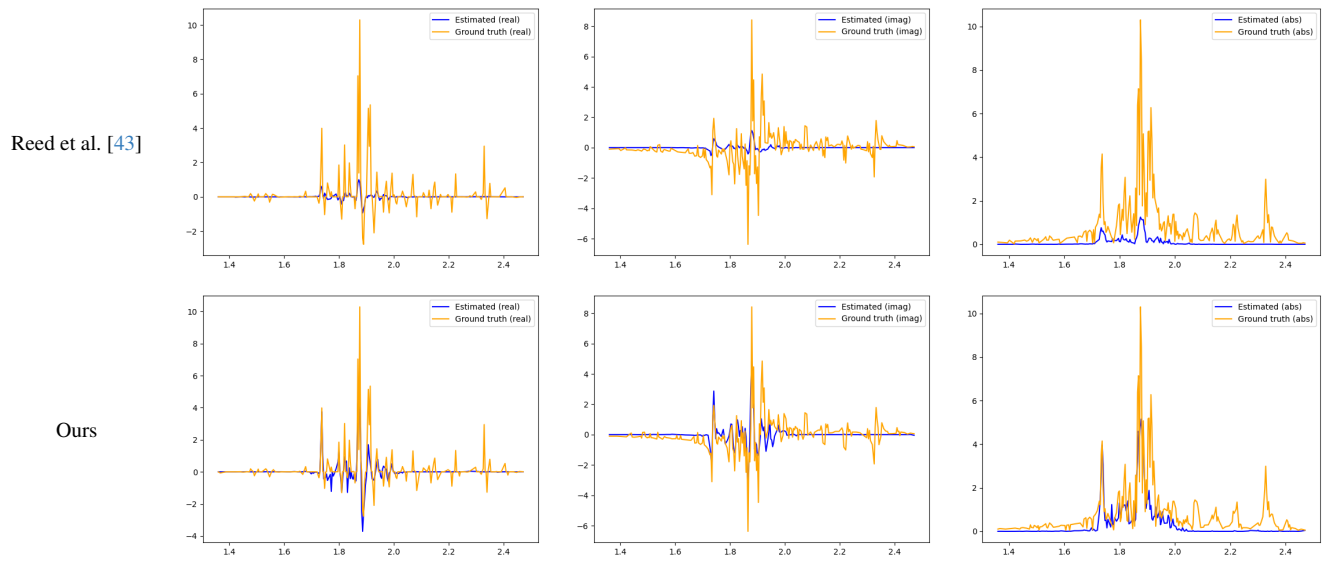


Figure 12. Novel view signal synthesis results comparing Reed et al. [43] (top row) and our method (bottom row). Each column shows a different component of the complex signal: real part (left), imaginary part (middle), and absolute magnitude (right). Within each plot, the estimated signal (blue) is compared against the ground-truth measurement (orange). Our method achieves closer alignment with ground-truth across all components, especially in regions with sharp variations.