Supplemental Material: Neural Volumetric Reconstruction for Coherent Synthetic Aperture Sonar

ALBERT REED, Arizona State University, USA
JUHYEON KIM, Dartmouth College, USA
THOMAS BLANFORD, The Pennsylvania State University, USA
ADITHYA PEDIREDLA, Dartmouth College, USA
DANIEL C. BROWN, The Pennsylvania State University, USA
SUREN JAYASURIYA, Arizona State University, USA

ACM Reference Format:
Albert Reed, Juhyeon Kim, Thomas Blanford, Adithya Pediredla, Daniel C. Brown, and Suren Jayasuriya. 2023. Supplemental Material: Neural Volumetric Reconstruction for Coherent Synthetic Aperture Sonar. *ACM Trans. Graph.* 42, 4 (August 2023), 9 pages. https://doi.org/10.1145/3592141

In this supplemental material, we present a derivation of our analytic forward model, additional results, and ablation studies to justify our method presented in the main paper. We also encourage the reader to view the supplemental videos of revolving 3D reconstructions as well as the code and data[1].

1 DERIVATION OF ANALYTIC FORWARD MODEL WITH PULSE DECONVOLVED MEASUREMENTS

We first provide the full derivation of our analytic forward model to synthesize pulse deconvolved waveforms. Our starting point is the forward model given by point-based sonar scattering [Brown 2017; Brown et al. 2017]:

$$\mathbf{s}(t) = \int_{\mathcal{X}} \frac{b_T(\mathbf{x}) b_R(\mathbf{x}) T(\mathbf{o}_T, \mathbf{x}) T(\mathbf{o}_R, \mathbf{x})}{2\pi R_T R_R} \sigma(\mathbf{x}) p\left(t - \frac{R_t + R_R}{c}\right) \mathrm{d}\mathbf{x}. \tag{1}$$

We redefine the terms in Eq. 1 for completeness. The real-valued sonar measurements in time are denoted by $\mathbf{s}(t)$. The set of all scene points is given by $\mathcal{X}$, and an individual point is given by $\mathbf{x}$. $\mathbf{o}_T$ and $\mathbf{o}_R$ denote the transmitter and receiver positions, respectively. The transmitter and receiver directivity functions are given by $b_T(\mathbf{x})$ and $b_R(\mathbf{x})$. The transmission probability from the transmitter to a scene point is given by $T(\mathbf{o}_T, \mathbf{x})$; its probability of return back to the receiver is given by $T(\mathbf{o}_R, \mathbf{x})$. The distance from the transmitter to a scene point is given by R_T; R_R denotes the distance from the point back to the receiver. The scene scatterer amplitudes are given by the function $\sigma(\cdot)$. The transmit pulse in time is defined by $p(t)$ and is by the time-of-flight to a scene point using the sound speed, c.

For the sake of readability, we define the term

$$K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) = \frac{b_T(\mathbf{x}) b_R(\mathbf{x}) T(\mathbf{o}_T, \mathbf{x}) T(\mathbf{o}_R, \mathbf{x})}{2\pi R_T R_R}$$

for the integral in this equation and subsequent derivations. We also note that σ can be any scattering model including the Lambertian model $L(\sigma)$ used in the main paper.

If we perform cross-correlation on the equation above, we would compute

$$\mathbf{s}(t) = \mathbf{s}(t) *_t p^*(-t). \tag{2}$$

Combining Eq. (1) and (2) and using the linearity properties of the convolution operator, we get the following equation:

$$\mathbf{s}(t) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \sigma(\mathbf{x}) \cdot p\left(t - \frac{R_t + R_R}{c}\right) *_t p^*(-t) \mathrm{d}\mathbf{x}. \tag{3}$$

Constructing the analytic signal for the measurement given in Eq. (3), we get that:

$$\widehat{\mathbf{s}} = \mathbf{s} + j\mathcal{H}(\mathbf{s}), \tag{4}$$

$$\widehat{\mathbf{s}} = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \sigma(\mathbf{x}) \cdot p\left(t - \frac{R_t + R_R}{c}\right) *_t p^*(-t) \mathrm{d}\mathbf{x} + j\mathcal{H}\left(\int_{\mathcal{X}} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \sigma(\mathbf{x}) \cdot p\left(t - \frac{R_t + R_R}{c}\right) *_t p^*(-t) \mathrm{d}\mathbf{x}\right). \tag{5}$$

Using the linearity of the Hilbert transform and regrouping terms,

$$\widehat{\mathbf{s}} = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \sigma(\mathbf{x}) \cdot \widehat{P}(t) \mathrm{d}\mathbf{x}, \tag{6}$$

where

$$\widehat{P}(t) = \left(p\left(t - \frac{R_t + R_R}{c}\right) *_t p^*(-t)\right) + j\mathcal{H}\left(p\left(t - \frac{R_t + R_R}{c}\right) *_t p^*(-t)\right).$$

Here, the term $\widehat{P}$ represents the analytic signal of the cross-correlated pulse.

[1] https://awreed.github.io/Neural-Volumetric-Reconstruction-for-Coherent-SAS/

Authors' addresses: Albert Reed, Arizona State University, USA, albertnm123@gmail.com; Juhyeon Kim, Dartmouth College, USA, juhyeon.kim.gr@dartmouth.edu; Thomas Blanford, The Pennsylvania State University, USA, teb217@psu.edu; Adithya Pediredla, Dartmouth College, USA, aditya.eee.nitw@gmail.com; Daniel C. Brown, The Pennsylvania State University, USA, dcb19@psu.edu; Suren Jayasuriya, Arizona State University, USA, sjayasur@asu.edu.

0730-0301/2023/8-ART $15.00
https://doi.org/10.1145/3592141

We now perform a common modeling trick of assuming the point scattering field is complex, i.e. $\widehat{\sigma'} = \sigma \cdot \widehat{P}$, and have our method estimate these values instead. Note that a similar trick was performed by Reed et al. [Reed et al. 2022] for 2D SAS image deconvolution.

$$\widehat{\mathbf{s}'_{\text{PD}}}\left(t = \frac{R_T + R_R}{c}\right) = \int_{\mathcal{X}} \frac{b_T(\mathbf{x})T(\mathbf{o}_T, \mathbf{x})T(\mathbf{x}, \mathbf{o}_R)}{2\pi R_T R_R} \widehat{\sigma'}(\mathbf{x})\mathrm{d}\mathbf{x}, \quad (7)$$

Ideal Pulse Deconvolution. If we assume that our pulse deconvolution described in Section 5 is ideal, then $p\left(t - \frac{R_t+R_R}{c}\right) *_t p^*(-t) = \delta\left(t - \frac{R_t+R_R}{c}\right)$ is an ideal delta function. Then we can simplify Eq. (6) as follows:

$$\widehat{\mathbf{s}}(t) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \sigma(\mathbf{x}) \cdot \widehat{\delta}\left(t - \frac{R_t + R_R}{c}\right) \mathrm{d}\mathbf{x}. \quad (8)$$

Computing the analytic signal of the delta function using the fact that $\mathcal{H}(\delta(t)) = \frac{1}{\pi t}$, we get:

$$\widehat{\delta}\left(t - \frac{R_t + R_R}{c}\right) = \delta\left(t - \frac{R_t + R_R}{c}\right) + \frac{j}{\pi\left(t - \frac{R_t+R_R}{c}\right)} \quad (9)$$

Now we can use the fact that $\widehat{P} = \widehat{\delta}\left(t - \frac{R_t+R_R}{c}\right)$ has most of its energy at the time-of-flight $t = \frac{R_t+R_R}{c}$. As shown in Fig. 5 in the main paper, if we assume a one-bounce reflection model, then the set of points with a constant time-of-flight $t = \frac{R_t+R_R}{c}$ describe an ellipsoid with a semi-major axis of length $r = c \cdot t/2$, where c is the sound speed. Specifically, these points define the ellipsoid:

$$\frac{x^2}{a(r)^2} + \frac{y^2}{b(r)^2} + \frac{z^2}{c(r)^2} - 1 = 0, \quad (10)$$

where transmit $\mathbf{o}_T$ and receive $\mathbf{o}_R$ elements are separated by distance d and the ellipsoid axes are,

$$a(r) = r, b(r) = \sqrt{(r)^2 - (d/2)^2}, c(r) = b(r). \quad (11)$$

We let $\mathbf{E}_r$ be the set of $\mathbf{x}$ points on the surface of the ellipsoid defined by range r. Thus, we can approximate Eq. (7) since Eq. (9) has most of its energy near $t = \frac{R_t+R_R}{c}$, and thus restrict the domain of integration to $\mathbf{E}_r$:

$$\widehat{\mathbf{s}}\left(t = \frac{R_t + R_R}{c}\right) \approx \int_{\mathbf{E}_r} K(\mathbf{x}, \mathbf{o}_R, \mathbf{o}_T) \cdot \widehat{\sigma'}(\mathbf{x})\mathrm{d}\mathbf{x}. \quad (12)$$

Equation (12) is the final forward model that we utilize in our main method to synthesize $\widehat{\mathbf{s}}(\mathbf{t})$ that is optimized against the real measurements from the sonar. In implementation, we omit the spherical spreading term $\frac{1}{2\pi R_T R_R}$ in the equation because its effects are negligeable for our relatively small scene sizes. We note that this term is commonly ommitted in time-domain beamformer implementations [Hayes and Gough 1992]. Limitations of the forward model include not modeling diffraction, a single bounce assumption, and difficulty recovering elastic scattering effects. We discuss these limitations in more detail in the main paper.

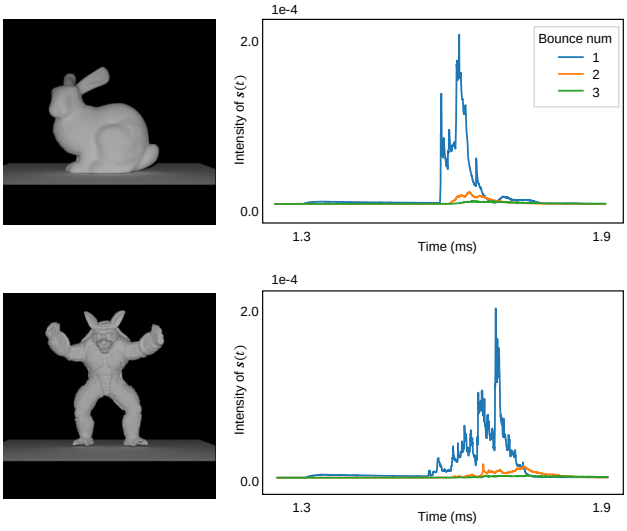

Fig. 1. Rendered image and transient signal from one sensor location for the bunny and armadillo objects.

2 TIME OF FLIGHT RENDERER DETAILS

In this section, we describe the details of the renderer implementation used to simulated SAS measurements. We use a transient renderer that is bootstrapped to the optical renderer [Kim and Kim 2021], which uses the GPU ray tracing library, OptiX. The transient renderer exploits a CUDA atomic operation on top of the optical renderer to render transient signals. We render each object at 54000 different camera poses, each pose corresponding to a measurement position from AirSAS, and the transient signal sampled at the AirSAS sampling frequency of 100 kHz.

Examples of rendered AirSAS scenes and the corresponding transient signals from one sensor location can be found in Fig. 1. We add a plane below the objects to emulate the physical AirSAS setup, which is confined with an anechoic chamber that is padded everywhere except the ground. Accumulating rendered rays by their bounce number shows that the intensity of the single-bounced signal is dominant.

3 ADDITIONAL SIMULATED RESULTS

We show the results for simulated data for both shape reconstruction (Fig. 2) and surface normal reconstruction (Fig. 3). All simulations assume $\Delta f = 20kHz$ and a signal-to-noise ratio of 20dB. The quantitative metrics from these eight meshes are summarized in Table 2 in the main paper. Our method can consistently achieve high-quality reconstructions more faithful to the ground truth meshes in both shape and surface normal distribution.

4 ABLATION STUDIES

Coherent versus incoherent reconstruction: In the main paper, we describe how we compute our loss between the complex-valued analytic measurements and network synthesized measurements. In particular, our network outputs a complex-valued scene $\widehat{\sigma'}$ to

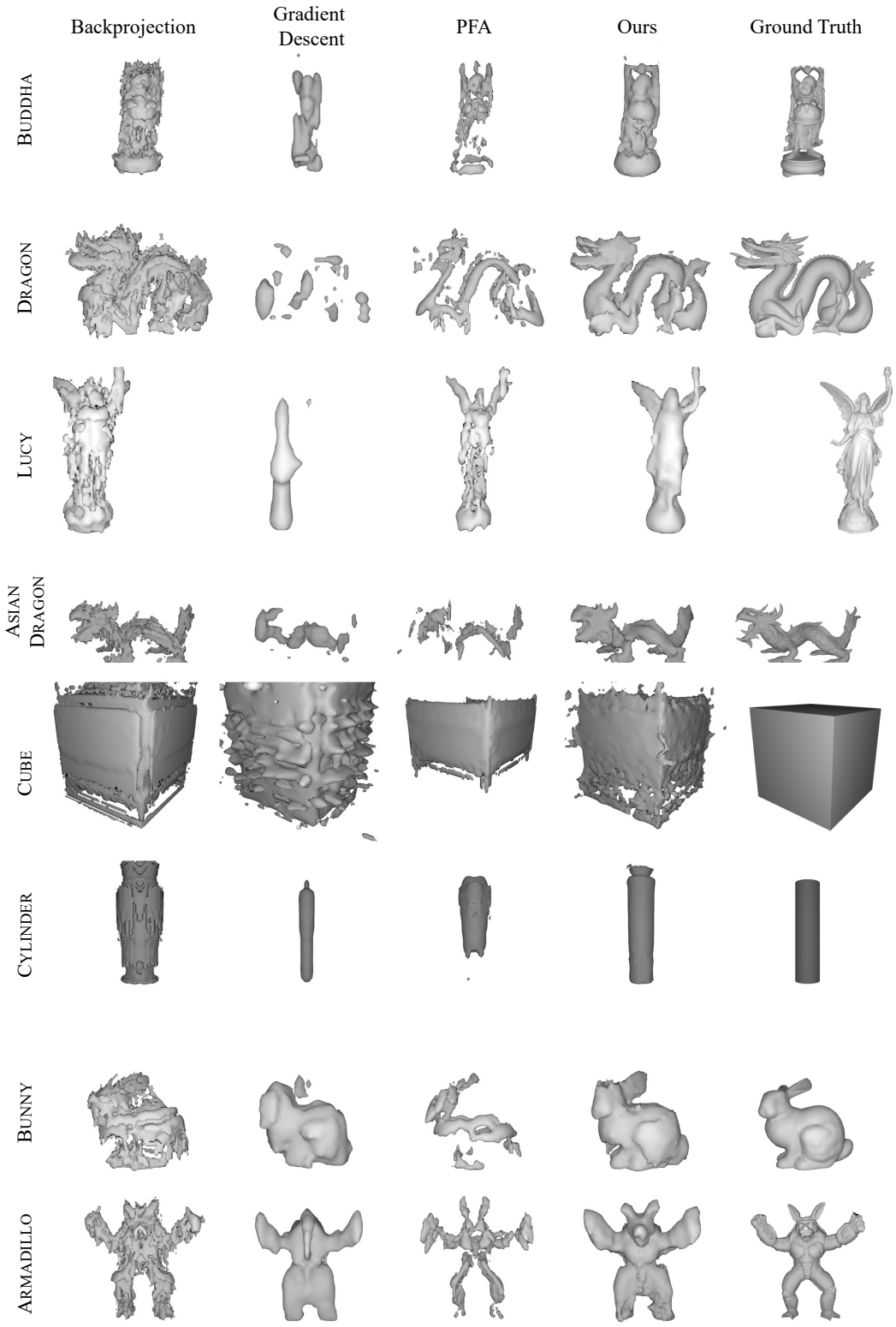

Fig. 2. Rendered meshes using different methods. The simulation was performed under bandwidth 20k with a noise level of 20db. We observe that our technique consistently performs better than traditional techniques.

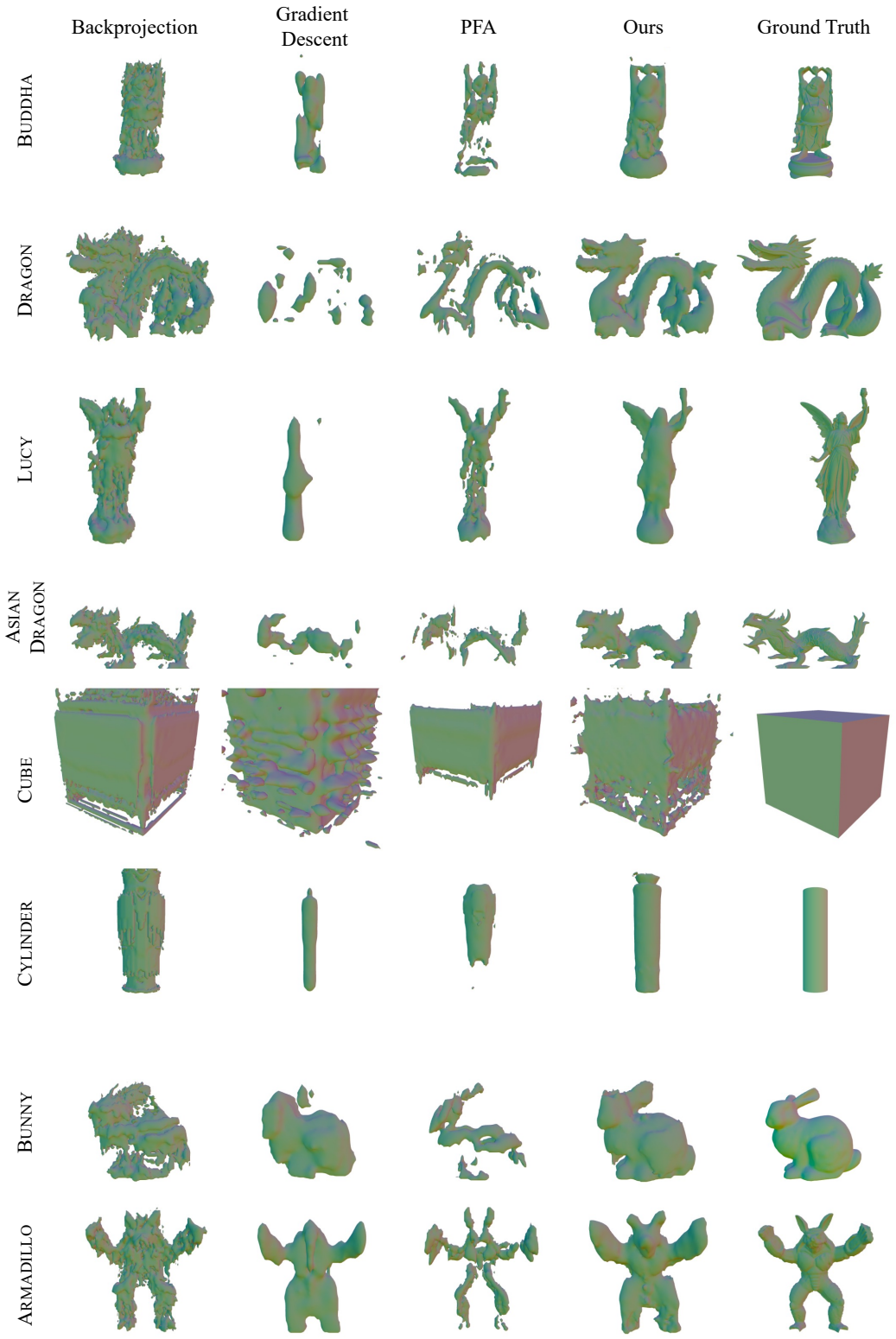

Fig. 3. Normal of rendered meshes using different methods. The simulation was performed under bandwidth 20k with the noise level of 20db. Our technique reconstructs normal maps that are more similar to the ground truth than previous techniques.

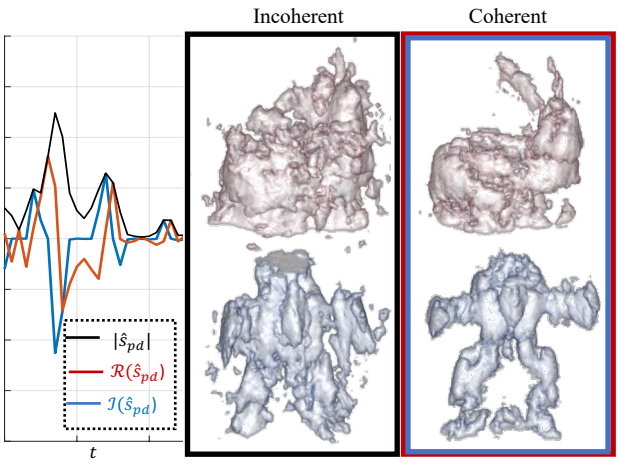

Fig. 4. Reconstructions from our method using incoherent and coherent processing of $\Delta f = 20$ kHz AirSAS bunny and armadillo. Incoherent processing integrates the envelope (black line) of measurements; coherent processing integrates the complex analytic signal (red and blue lines) of measurements. Clearly, coherent processing is necessary for maximizing performance.

synthesize these complex measurements. In contrast, some sonar reconstructions (e.g. forward-look) leverage an incoherent reconstruction where only the magnitude of the signal is used [Kim 2007; Kim et al. 2005]. To compare coherent versus incoherent reconstruction, we configured our network to output a real-valued scene to synthesize the magnitude of measurements directly for incoherent reconstruction, which we show in Fig. 4. The left column of the figure shows the complex analytic signal (red and blue curves) used in the coherent reconstruction and their magnitude (black curve) which is used for the incoherent reconstruction. The performance disparity aligns with SAS processing fundamentals that measurements should be coherently processed to enhance resolution.

Effect of regularization terms: Finally, we studied the impact of our scene regularizations in Fig. 5. The neural backprojection smoothness and sparsity priors are sometimes useful in certain conditions. We find that they are especially helpful for sparse view reconstructions, as they attenuate streaking artifacts that occur due to having limited angular measurements. In the figure, we show that using the smoothness and sparsity priors attenuates noise in the scene and enables a more accurate reconstruction of the armadillo object.

5 ADDITIONAL SVSS RESULTS

SVSS Quantitative: In Figure 6, we show the maximum intensity projection (MIP) in the depth dimension for the same cinder block targets shown in the main text. While measuring the dimensions of the targets from the MIP is an approximation since we are actually measuring the projection of pixels onto a plane, we show that our method retains most of the accuracy of backprojection, typically reconstructing the target dimensions typically within a few centimeters of ground truth. Our method tends to estimate a size slightly smaller than ground truth, whereas backprojection

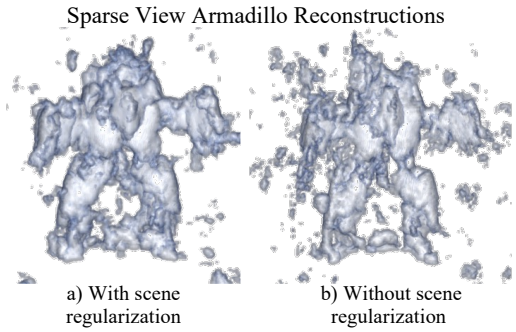

Fig. 5. Effect of Prior: (Ablations using sparse view AirSAS measurements of the 20 kHz armadillo). We show a sparse sampled armadillo reconstruction with and without scene priors. Scene priors are especially helpful in the sparse and helical sampling cases to attenuate undersampling artifacts.

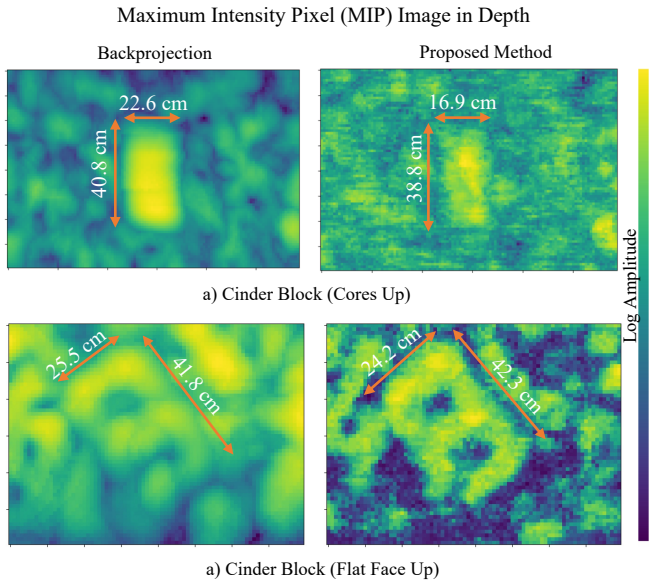

Fig. 6. MIP images of the (a) cinder block with cores up and (b) cinder block face up. The ground truth dimensions for the cinder block are 20×40 centimeters (cm). Our proposed method performs comparably to backprojection in the accuracy of the reconstructed dimensions. Note that each image pixel is 1×1cm.

estimates slightly larger. This is perhaps because of our pulse deconvolution step compressing measurements in range more than matched filtering.

6 ADDITIONAL AIRSAS RESULTS

Effects of rendering threshold parameter: In the main text, we normalize the magnitude of AirSAS reconstructions between $[0 - 1]$ and use a threshold of 0.2 for visualization (i.e. set all magnitudes $< 0.2 = 0$). In Fig. 7 and Fig. 8, we render the AirSAS results from the main text at higher (0.3) and lower (0.05) thresholds, respectively. While the threshold affects the qualitative appearance of all

methods, we observe that our method retains better retains object geometry across all thresholds when compared with backprojection and gradient descent.

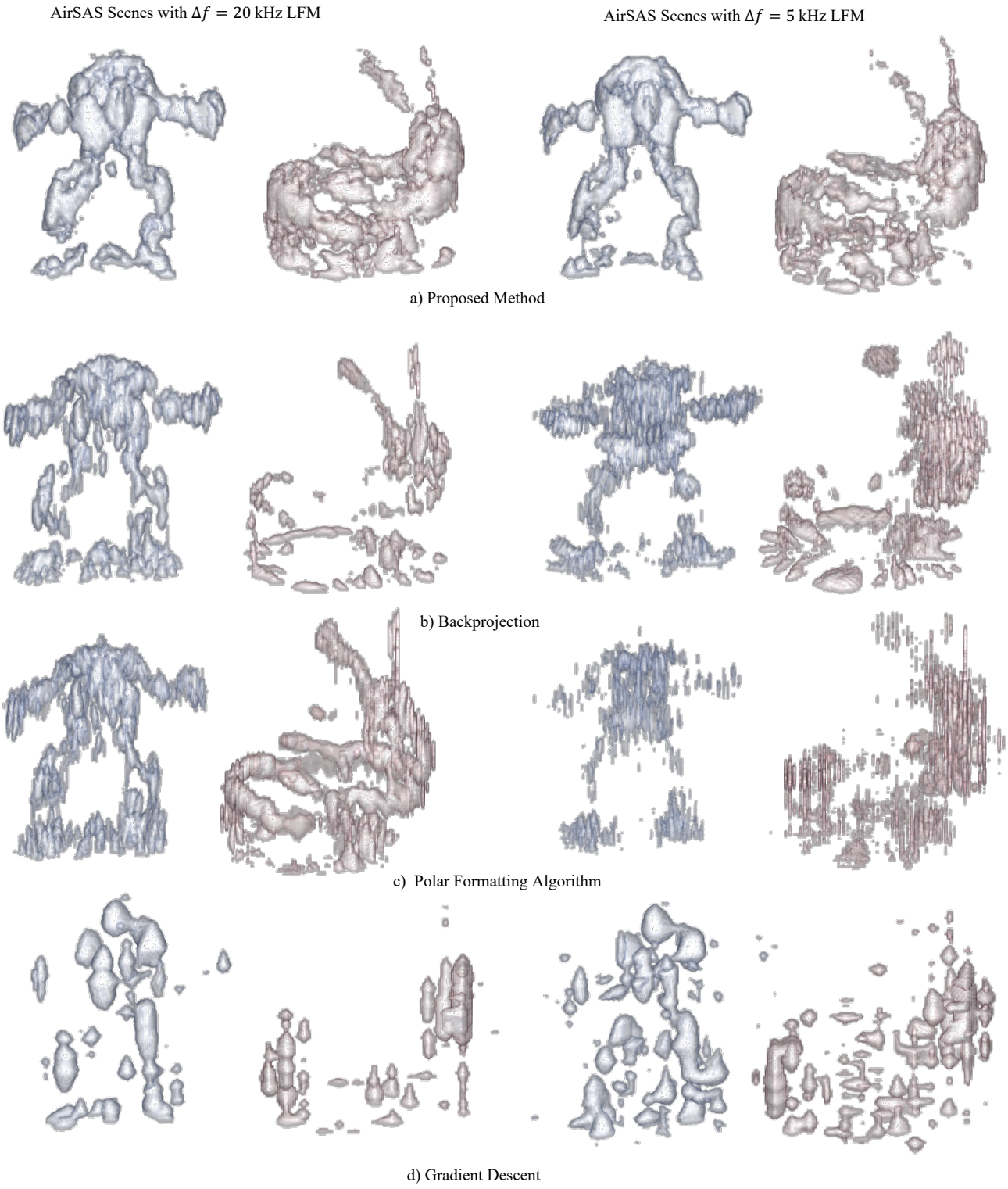

Fig. 7. Higher threshold used to visualize reconstructions of AirSAS data captured with relatively high ($\Delta f = 20$ kHz) and low ($\Delta f = 5$ kHz) bandwidth LFMs. Our method demonstrates more consistent performance across waveform bandwidth compared to backprojection, the polar formatting algorithm, and gradient descent.

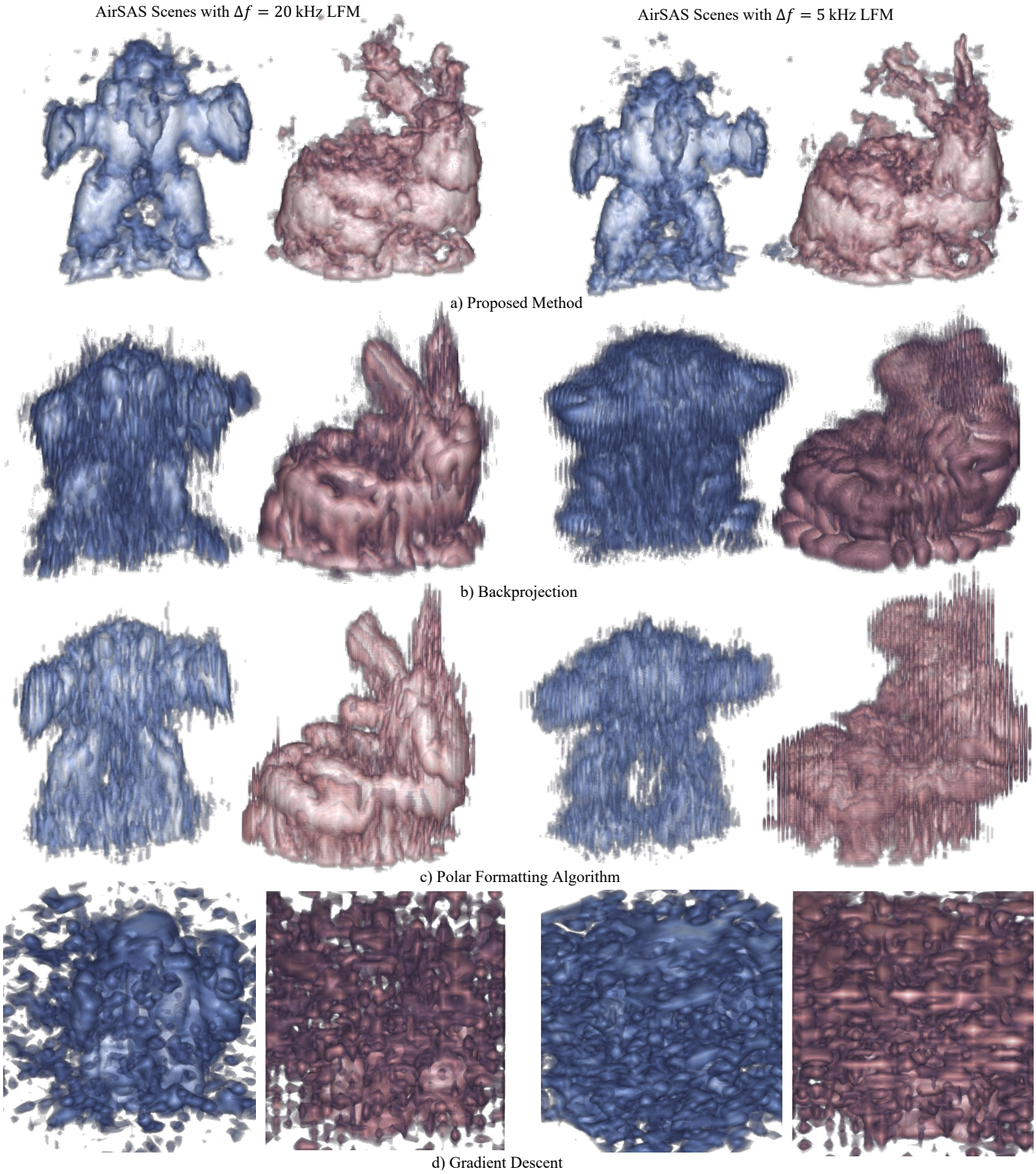

Fig. 8. Lower threshold used to visualize reconstructions of AirSAS data captured with relatively high ($\Delta f = 20$ kHz) and low ($\Delta f = 5$ kHz) bandwidth LFMs. Our method demonstrates more consistent performance across waveform bandwidth compared to backprojection, the polar formatting algorithm, and gradient descent.

REFERENCES

Daniel C Brown. 2017. *Modeling and measurement of spatial coherence for normal incidence seafloor scattering.* Ph. D. Dissertation. The Pennsylvania State University.

Daniel C Brown, Shawn F Johnson, and Derek R Olson. 2017. A point-based scattering model for the incoherent component of the scattered field. *The Journal of the Acoustical Society of America* 141, 3 (2017), EL210–EL215.

Michael P Hayes and Peter T Gough. 1992. Broad-band synthetic aperture sonar. *IEEE Journal of Oceanic Engineering* 17, 1 (1992), 80–94.

Juhyeon Kim and Young Min Kim. 2021. Fast and Lightweight Path Guiding Algorithm on GPU. In *Pacific Graphics Short Papers, Posters, and Work-in-Progress Papers*, Sung-Hee Lee, Stefanie Zollmann, Makoto Okabe, and Burkhard Wünsche (Eds.). The Eurographics Association. https://doi.org/10.2312/pg.20211379

Kio Kim. 2007. *Enhanced echolocation via robust statistics and super-resolution of sonar images.*

Kio Kim, Nicola Neretti, and Nathan Intrator. 2005. Mosaicing of acoustic camera images. *IEE Proceedings-Radar, Sonar and Navigation* 152, 4 (2005), 263–270.

Albert Reed, Thomas Blanford, Daniel C. Brown, and Suren Jayasuriya. 2022. SINR: Deconvolving Circular SAS Images Using Implicit Neural Representations. *IEEE Journal of Selected Topics in Signal Processing* (2022), 1–16. https://doi.org/10.1109/JSTSP.2022.3215849