Analyzing Time of Flight Radiance Fields for Multi-Frequency Reconstruction

RUNFENG LI, Brown University, USA
JACK NAYLOR, The University of Sydney, Australia
MIKHAIL OKUNEV, Brown University, USA
CHRISTIAN RICHARDT, Meta Reality Lab, Switzerland
MATTHEW O'TOOLE, Carnegie Mellon University, USA
JAMES TOMPKIN, Brown University, USA

Continuous-wave time-of-flight (C-ToF) cameras are compact, inexpensive sensors that deliver dense per-pixel depth at video frame rates, making them well-suited for 3D scene reconstruction. Fitting a volumetric radiance field to raw C-ToF phasor measurements is ill-posed: many density distributions along a ray produce the same measured phasor. One unexplored approach to overcome this is with multiple frequencies, but this too is ill posed. We investigate how to make this work through multiple steps. First, by reassessing how density-noise regularizations in prior single-frequency work encourage meaningful scene depths, with analysis through the spread of density in phasor sums. Second, through reconsidering the Cartesian L2 phasor loss, which fails to accommodate large amplitude variation in multi-frequency settings. We replace it with an amplitude-normalized phasor loss that approximates the log-polar error metric, correctly weighting phase and relative-amplitude errors across the full measurement dynamic range rather than penalizing dim and bright returns equally. Third, we extend this approach to multiple simultaneous modulation frequencies from two views: the same scene density drives renderings at every frequency, so the cross-frequency coupling of rendered phasors provides additional constraints that reduce depth ambiguity, enable principled phase unwrapping, and handle cross-sensor interference in multi-camera configurations. Our method produces higher-quality 3D reconstructions than prior C-ToF radiance-field methods, recovering thin structures and low-reflectance objects that past approaches fail to capture.

Additional Key Words and Phrases: time-of-flight imaging, radiance fields, depth estimation, multi-frequency, phasor, scene reconstruction

ACM Reference Format:
Runfeng Li, Jack Naylor, Mikhail Okunev, Christian Richardt, Matthew O'Toole, and James Tompkin. 2026. Analyzing Time of Flight Radiance Fields for Multi-Frequency Reconstruction. In *Proceedings of XXXXXX (XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXX.XXXXXX

Authors' Contact Information: Runfeng Li, runfeng_li@brown.edu, Brown University, USA; Jack Naylor, jack.naylor@sydney.edu.au, The University of Sydney, Australia; Mikhail Okunev, mikhail_okunev@brown.edu, Brown University, USA; Christian Richardt, christian@richardt.name, Meta Reality Lab, Switzerland; Matthew O'Toole, , Carnegie Mellon University, USA; James Tompkin, james_tompkin@brown.edu, Brown University, USA.

XX, XXX

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXX.XXXXXX

1 Introduction

Continuous-wave time-of-flight (C-ToF) cameras produce dense per-pixel depth from inexpensive sensors by recovering the phase of returned light, with per-pixel depth bound to an unambiguous range $c/(2f)$ set by the modulation frequency f. Using multiple frequencies lets us extend the range by 'unwrapping' the phases in closed form via a GCD relation [10], and even reconstruct the scene behind translucent materials by modeling multi-path effects [4, 11]. Recently, a separate line of work models C-ToF within a volumetric neural radiance field [1, 5, 23, 31] based on raw phasor measurements [12]. Lifting samples into 3D with more principled modeling offers denoising and multi-sensor integration opportunities.

Yet, even with this model, the fitting problem remains ill-posed: many density distributions along a ray produce the same rendered phasor, and only some yield a faithful rendered depth. Why do these methods work at all, and what does multi-frequency change? To answer this, we analyze the rendered phasor sum on the complex plane to expose the equivalence class of density distributions that the typical existing loss admits. We show that an angular coupling between frequencies reduces spread densities but does not break the equivalence class. Empirically, we show how a critical density-noise heuristic helps recover a useful distribution in single- and multi-frequency cases. Then, we show that the standard Cartesian phasor loss inadvertently penalizes the same absolute error equally for dim and bright pixels, and that replacing it with an amplitude-normalized loss whose iso-loss radius scales with phasor magnitude is critical when return amplitudes vary widely across the image.

Adding a second simultaneous camera may also help. We derive a multi-view multi-frequency method in which cross-frequency leakage between simultaneously-firing emitters from two cameras is sinc suppressed. Given that, we can exploit the complementary geometric constraint from two views on the shared 3D field. The multi-view constraint has been carrying much of the empirical success of prior single-frequency methods, with density-noise regularization doing more of the work in sparse-view or single-view cases. Finally, with our complete analysis across multi-frequency and multi-view settings, we can capture and optimize to recover 3D scenes beyond the single-frequency unambiguous range, with no interference, with low noise, and with fine reconstruction of small features. And, as a side benefit, we can resolve translucent scenes to measure depths through objects.

Assumptions and Limitations. We restrict ourselves to static scenes to simplify our analysis. We adopt the standard sinusoidal C-ToF

idealization and treat sensor demodulation contrast as a single per-frequency scalar, ignoring per-pixel and amplitude non-linearities. We do not explicitly model reflection, refraction, or scattering. Reconstruction is iteratively optimized per scene and is not real-time. We primarily use one lab scene across different capture variants, which shows the discovered effects in a consistent way.

2 C-ToF Camera Principles

C-ToF cameras measure depth by emitting sinusoidally modulated illumination at frequency f and correlating the returned light against phase-shifted references [14]. Under a *single-return model*, in which one static surface returns light during the four quad exposures, the light arriving at each pixel is

$$L_r(t) \;=\; B_0 + A\sin(2\pi f t - \psi), \tag{1}$$

with a DC bias B_0 (ambient light and sensor offset) plus a sinusoid of amplitude A and round-trip phase $\psi = 4\pi f d/c$ set by the surface depth d. Each quad time-averages $L_r(t)$ against a phase-shifted reference at the same frequency,

$$Q_\phi \;=\; \frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} 2\sin(2\pi f t - \phi)\, L_r(t)\,\mathrm{d}t, \tag{2}$$

with exposure time $\mathcal{T}$ and offsets $\phi \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$; the factor of two normalizes the reference so that the recovered amplitude matches A in Equation (1). After product-to-sum reduction, the oscillating components in the integrand average to zero for $\mathcal{T} \gg 1/f$, leaving the constant

$$Q_\phi = A\cos(\psi - \phi). \tag{3}$$

Three offsets suffice in principle to recover the three unknowns (A, ψ, B_0); using four equispaced offsets cancels B_0 algebraically through opposite-phase differences,

$$Q_0 - Q_\pi = 2A\cos\psi, \qquad Q_{\frac{3\pi}{2}} - Q_{\frac{\pi}{2}} = 2A\sin\psi, \tag{4}$$

giving phase by four-quadrant arctangent

$$\psi = \operatorname{atan2}\Big(Q_{\frac{3\pi}{2}} - Q_{\frac{\pi}{2}},\; Q_0 - Q_\pi\Big) \tag{5}$$

and amplitude as the magnitude of the bias-canceled signal,

$$A = \tfrac{1}{2}\sqrt{(Q_0 - Q_\pi)^2 + (Q_{\frac{3\pi}{2}} - Q_{\frac{\pi}{2}})^2}. \tag{6}$$

Equivalently, the four quads define a complex phasor

$$p = \tfrac{1}{2}\big[(Q_0 - Q_\pi) \;-\; j\,(Q_{\frac{\pi}{2}} - Q_{\frac{3\pi}{2}})\big] = Ae^{j\psi}, \tag{7}$$

whose magnitude is the returned amplitude and whose argument ψ encodes the round-trip light travel time, converting to depth as

$$d \;=\; \frac{c}{4\pi f}\,\psi, \tag{8}$$

with c the speed of light. Because $\psi \in [0, 2\pi)$ is periodic, a single frequency recovers depth only within the unambiguous range

$$d_{\max} = \frac{c}{2f}. \tag{9}$$

A second sensor non-ideality is the *demodulation contrast* $\eta(f) \in (0, 1]$, an efficiency factor that attenuates the reported amplitude relative to the ideal correlator output and decreases with f across the usable band. We treat it as a per-frequency scalar gain on A, calibrated or estimated as a ratio across frequencies from the median amplitude.

Multi-frequency capture extends the unambiguous range. Capturing at multiple modulation frequencies $\{f_k\}$ and combining the wrapped phases ψ_k recovers depth from the unwrapped phase Ψ as

$$d = \frac{c}{4\pi\,\mathrm{GCD}(f_1, \ldots, f_K)}\,\Psi, \tag{10}$$

where GCD is the greatest common divisor of the chosen frequencies [10]. A small GCD yields a long unambiguous range; for example, $f_1 = 20$ MHz and $f_2 = 30$ MHz (GCD = 10 MHz) extend $d_{\max}$ from 7.5 m to 15 m, with the practical limit set by the $1/d^2$ signal falloff.

Residual failure modes. Even with multi-frequency capture, the single-return model still fails in low light, at high noise, or when multiple paths or semi-transparent layers contribute to one pixel [4, 11]. These failures motivate a physically-based forward model fit directly to the raw C-ToF measurements; as later sections show, achieving this is harder than the additional multi-frequency constraints might suggest.

3 Analyzing ToF Radiance Fields

ToF radiance fields optimize a volumetric scene representation directly against raw C-ToF measurements rather than against camera-derived depth maps. Prior work has shown that this is feasible even from a single fixed camera, despite the image-formation model being ill-posed. We ask why (and when) it works, and connect the mechanism we identify to the multi-frequency and multi-view settings analyzed in the following sections.

3.1 Forward model

Consider a camera ray from camera center $\mathbf{x}$ in direction $\boldsymbol{\omega}_o$, with samples $\mathbf{x}_s = \mathbf{x} + s\boldsymbol{\omega}_o$. A ToF radiance field represents density $\sigma(\mathbf{x}_s)$ and returned active-light amplitude $L_a(\mathbf{x}_s, \boldsymbol{\omega}_o)$ from light that travels from the source to the scene and back. Assuming the emitter is collocated with the sensor, each sample is weighted by squared transmittance and inverse-square falloff. For modulation frequency f, the rendered phasor is

$$\hat{p}_f(\mathbf{x}, \boldsymbol{\omega}_o) = \int_{s_n}^{s_f} \frac{T(\mathbf{x}, \mathbf{x}_s)^2}{d_s^2}\sigma(\mathbf{x}_s)L_a(\mathbf{x}_s, \boldsymbol{\omega}_o)\exp\left(j\frac{4\pi f d_s}{c}\right)\mathrm{d}s, \tag{11}$$

where $d_s = \|\mathbf{x} - \mathbf{x}_s\|$, T is transmittance, σ is volume density, and c is the speed of light.

Equivalently, one can render the four raw C-ToF quads with phase offsets $\phi \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$:

$$Q_\phi(d_s, f) = \cos\left(\frac{4\pi f d_s}{c} - \phi\right). \tag{12}$$

The corresponding rendered quad is

$$\hat{q}_{\phi,f}(\mathbf{x}, \boldsymbol{\omega}_o) = \int_{s_n}^{s_f} \frac{T(\mathbf{x}, \mathbf{x}_s)^2}{d_s^2}\sigma(\mathbf{x}_s)L_a(\mathbf{x}_s, \boldsymbol{\omega}_o)Q_\phi(d_s, f)\mathrm{d}s. \tag{13}$$

For a static scene over the four quad exposures, the quads contain the same phase and amplitude information as the corresponding phasor (Equations (5) and (6)); rendering raw quads at different times provides opportunities to handle dynamic scenes [23, 31].

Fig. 1. **Density noise drives geometry recovery.** Two-frequency single-view ToF radiance field reconstruction (42/55 MHz) for density noise standard deviations $s \in \{0.0, 0.3, 1.0, 1.5, 2.5, 3.5\}$ (left to right). **Top row:** rendered mean depth. **Bottom row:** rendered amplitude. Even with two frequencies, geometry only emerges at $s \geq 2.5$; the rendered amplitudes (bottom row) show that *phasor reconstruction is accurate in every case*, so density noise selects within a phasor-equivalent family. 50,000 iterations, learning rate 10^{-4}.

A standard supervision loss can be written either in phasor form or in raw quad form,

$$\mathcal{L}_{\mathrm{p}} = \sum_{\text{rays}} \left\| \hat{p}_f - p_f \right\|_2^2, \qquad \mathcal{L}_{\mathrm{q}} = \sum_{\text{rays}} \sum_{\phi} \left\| \hat{q}_{\phi,f} - q_{\phi,f} \right\|_2^2, \tag{14}$$

which we revisit in Section 3.3. Optimization dynamics differ between the two formulations, but the choice is independent of the scene representation. Together with these losses, Equations (11) and (13) let us optimize a radiance field directly against the raw C-ToF measurements, bypassing the camera's depth pipeline and the failure modes that affect it (Section 2). However, this does not imply that the recovered radiance field describes the scene geometry we want. For the analysis that explains why, we use phasor notation; the adaptation to raw quads follows naturally.

3.2 Phasor-fitting view and depth ambiguity

The forward model in Equation (11) can be written as a sum of per-sample phasors along each ray, each with its own phase θ_i (the per-sample analogue of the camera-side phase ψ in Equation (5)):

$$\hat{p}_f \;=\; \sum_i w_i \, \exp\left(j \frac{4\pi f d_i}{c} \right) \;=\; \sum_i w_i \, e^{j\theta_i}, \tag{15}$$

where d_i is the sample depth and $w_i \geq 0$ collects per-sample opacity, returned active-light amplitude, transmittance, and falloff.

Equation (15) lets us see that the loss admits an equivalence class of density distributions, only some of which are useful. A true surface could be represented by a *concentrated* density at a single depth or a *spread* density through constructive and destructive phasor addition; only the concentrated density produces an accurate mean depth [23]. Figure 2 shows members of this class for the same target, *each of which* achieves high-quality phasor reconstruction, and the supplementary document derives how rendered mean depth diverges from the true depth as the spread increases. Yet, even in the single-view setting, TöRF often finds good solutions despite the ill-posedness.

Why high density noise favors concentrated solutions. The optimizer needs a bias to select preferred (concentrated) members of the equivalence class. For MLP backbones, TöRF inherits NeRF's noise-injection regularization [1, 27]: at each sample i, the raw density predicted by

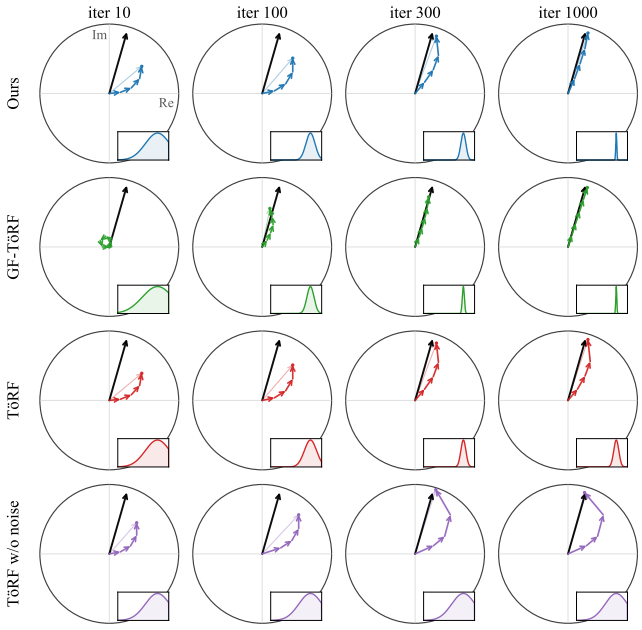


Fig. 2. **Phasor fitting across methods.** For a representative ray, the black arrow is the target phasor, the colored chained arrows are per-sample phasor contributions, and the faint arrow is their sum. Insets show the per-sample weight distribution along depth. Many distributions of contributions sum to the same target, but only weights concentrated at the correct depth yield meaningful geometry. Our method concentrates sharply at the correct depth; TöRF [1] sharpens only partially, and TöRF without density noise stays diffuse; GF-TöRF [23] concentrates by aligning phasor angles but fails beyond the unambiguous range, where these angles become ambiguous.

the network, σ_i^{raw}, is perturbed by additive Gaussian noise before the ReLU that enforces non-negativity,

$$\sigma_i \;=\; \max\!\left(0,\, \sigma_i^{\mathrm{raw}} + \epsilon_i\right), \qquad \epsilon_i \sim \mathcal{N}(0, s^2), \tag{16}$$

with ϵ_i independent across training samples. The noise is empirically essential: removing it preserves a faithful raw-phasor fit but produces poor rendered mean depth (Figure 1).

With larger s, the noise more often makes $\sigma^{\text{raw}} + \epsilon$ positive. The ReLU keeps it, so even samples where the network outputs zero or negative still receive some density on average. At $\sigma_i^{\text{raw}}=0$, $\mathbb{E}[\sigma_i] = s/\sqrt{2\pi}$, which is ≈ 1.2 at $s = 3$ and ≈ 2.0 at $s = 5$. These densities accumulate in the cumulative round-trip transmittance $T_i^2 = \exp\left(-2\sum_{k<i}\sigma_k\delta_k\right)$, so T^2 decays to near zero within a small front window of samples regardless of any phasor loss contribution.

The rendered phasor is then determined almost entirely by that first-hit window, with phase set by the window's depth. To match the captured phasor, the optimizer decreases σ^{raw} to be *confidently negative* at every depth where the surface should not be, deep enough that $\Pr\left(\sigma_i^{\text{raw}} + \epsilon_i > 0\right)$ stays small. As s increases (Figure 1), the noise lets the surface form density closer to the camera than at $s=0$. Past $s\approx 3$, there is little improvement.

This destabilization acts positionally: front samples are suppressed first, exposing deeper samples to the same pressure, and the resulting 'front-to-back sweep' settles wherever the phasor loss is locally minimized. In the single-camera setting without multi-view cues, for scenes with mostly empty space and opaque surfaces, this is a useful mechanism. Depth-distortion regularizers (Mip-NeRF 360 style) added to the phasor loss do not induce this sweep: they penalize spread weight distributions and so produce a more peaked $w(t)$ for any phasor-fit minimum, but the peak forms wherever the optimizer was already settling; distortion regularizes the *shape* of the solution, not its *position* along the ray.

Geometric reading: coherence. We can read the mechanism on the phasor circle (Figure 2). Each ray sample i contributes an arrow $w_i e^{j\theta_i}$ with direction $\theta_i = 4\pi f d_i/c$ set by the sample's depth and length $w_i = T_i^2\alpha_i L_a/d_i^2$ set by its weight. Summed tip-to-tail, these arrows trace an *arc* on the complex plane whose endpoint is the rendered phasor $\hat{p}_f$ (Equation (15)). The arc's magnitude is $|\hat{p}_f| = C\sum_i w_i$, where the *coherence*

$$C \equiv \frac{\left|\sum_i w_i e^{j\theta_i}\right|}{\sum_i w_i} \in [0,1] \tag{17}$$

measures arrow alignment: $C = 1$ for arrows pointing one way, $C \to 0$ for arrows that fan out and partially cancel.

Optimization walks $\hat{p}_f$ toward the captured p_f through arc reshapes. Without noise, the arc represents a low-coherence member of the spread/concentrated equivalence class: many arrows of moderate length spanning a wide range of θ_i, the arc curving through the disk on its way to p_f. Density noise reshapes this through the front-T^2 mechanism: front samples are suppressed to $\sigma \approx 0$, density peaks near the surface, and behind it T^2 has decayed so σ no longer matters. Tip-to-tail, the arrows are tiny at the front, lurch in one direction at the surface, and tiny again after: a near-straight, high-coherence arc.

3.3 Amplitude-normalized phasor loss

Density noise reshapes the optimization dynamics to prefer sparse density distributions; next, we consider the objective term itself. The Cartesian L2 losses $\mathcal{L}_{\text{p}}$ and $\mathcal{L}_{\text{q}}$ defined above treat all phasor errors equally on the complex plane, which is misaligned with what we care about in a phasor measurement: what matters is the phase and *relative* amplitude, not the absolute Cartesian distance from the target.

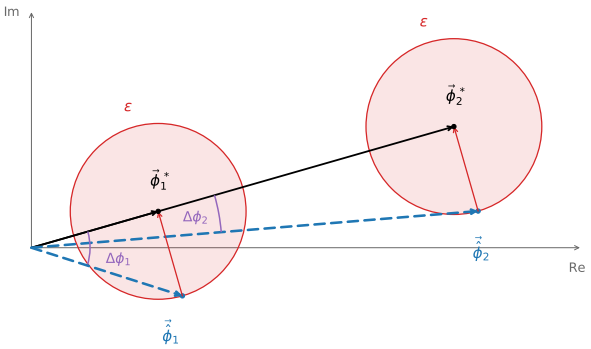
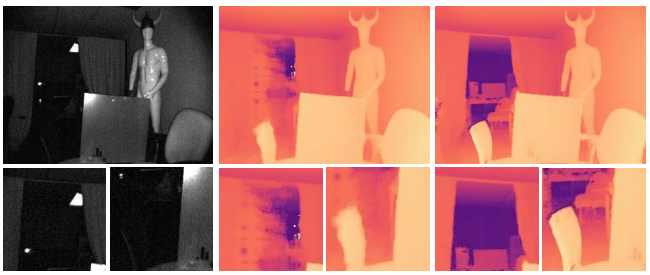


Fig. 3. **Geometric motivation for the amplitude-normalized phasor loss.** Two ground-truth phasors share a direction but differ in amplitude: $\vec{\phi}_1^*$ (dim pixel) and $\vec{\phi}_2^*$ (bright pixel). The predictions $\vec{\hat{\phi}}_1$ and $\vec{\hat{\phi}}_2$ sit at the *same* Cartesian distance ε from their targets, so an L_2 loss treats them as equally bad; yet the angular errors differ sharply, $\Delta\phi_1 \gg \Delta\phi_2$. Dividing by $|\vec{\hat{\phi}}|$ (Equation (23)) penalizes errors in proportion to $1/|\vec{\hat{\phi}}|$, which is large for dim phasors where the same Cartesian error implies a much larger phase error.



(a) C-ToF amplitude image (b) Optimized w/ L2 (c) Optimized w/ our loss

Fig. 4. **Amplitude-normalized phasor loss improves quality across the dynamic range.** C-ToF amplitudes vary widely across the scene, especially with multiple frequencies. A standard L_2 loss fits high-amplitude regions well but fails to recover depth in low-amplitude regions; our amplitude-normalized loss recovers both bright and dark regions.

We denote the predicted and captured phasors by

$$\hat{\mathbf{p}} = (\hat{x}, \hat{y}), \qquad \mathbf{p} = (x, y), \tag{18}$$

where x and y correspond to the cosine and sine components, respectively. The standard Cartesian loss

$$\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 = (\hat{x} - x)^2 + (\hat{y} - y)^2 \tag{19}$$

treats the same absolute Cartesian error equally everywhere in the complex plane. Figure 3 illustrates the issue: predictions at the same Cartesian distance ε from a dim and a bright ground-truth phasor have identical L_2 losses, yet the dim pixel's phase error $\Delta\phi$ is far larger than the bright pixel's, because the same Cartesian error produces a larger angular error when the phasor magnitude is small. More generally, for a small-magnitude phasor, the same Cartesian perturbation induces a larger phase error and a larger *relative* amplitude error than for a large-magnitude phasor.

We instead view each phasor in polar form,

$$r = \sqrt{x^2 + y^2}, \qquad \phi = \operatorname{atan2}(y, x), \tag{20}$$

and consider the more meaningful error variables

$$\Delta \log r = \log \hat{r} - \log r, \qquad \Delta\phi = \hat{\phi} - \phi. \tag{21}$$

The first measures relative amplitude error; the second measures phase error. In the spirit of RawNeRF's weighted loss [26], we linearize the transformed error around the prediction. For small perturbations, the log-polar error is locally approximated by

$$(\Delta \log r)^2 + (\Delta\phi)^2 \;\approx\; \frac{(\hat{x}-x)^2 + (\hat{y}-y)^2}{\hat{r}^2}. \tag{22}$$

This shows that the same absolute Cartesian error should be penalized more strongly for low-amplitude phasors than for high-amplitude ones. The full derivation is given in the supplementary document.

From this insight, we define the amplitude-normalized loss

$$\mathcal{L}_{\text{phasor}} = \frac{(\hat{x}-x)^2 + (\hat{y}-y)^2}{\text{sg}(\hat{x}^2 + \hat{y}^2) + \epsilon}, \tag{23}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and ϵ is a small constant for numerical stability. Without $\text{sg}(\cdot)$, the gradient would divide by the square of the predicted amplitude, potentially destabilizing training. Geometrically, Equation (23) preserves circular iso-loss contours in Cartesian phasor space but scales their radius with phasor magnitude: large-magnitude phasors are allowed proportionally larger absolute Cartesian errors, whereas small-magnitude phasors are fitted more strictly. Although this weighting is only a local approximation to the exact log-polar error, it matches the desired behavior in the regime relevant for optimization and is simple and stable in practice.

3.4 Demonstration: Improved Dark Regions

In Figure 4, we compare the amplitude-normalized loss $\mathcal{L}_{\text{phasor}}$ with $\epsilon = 0.01$ (Equation (23)) to the unweighted Cartesian L2 (Equation (19)), and optimize with Adam with learning rate 3×10^{-4}, batch size 1024, for 50,000 iterations. This loss provides a significant improvement to the reconstruction of both bright and dark areas of the scene.

4 Multi-frequency ToF Radiance Fields

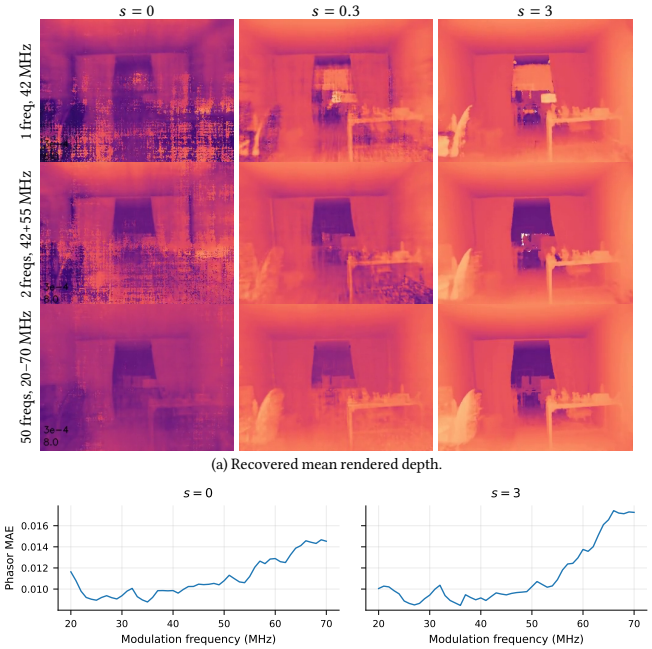
We now extend the radiance-field forward model to multi-frequency capture, one frequency at a time, and analyze what the additional measurements offer for the optimization.

4.1 Multi-frequency forward model

The single-frequency rendered phasor of Equation (11) extends to multiple frequencies by recognizing that the same scene density $\sigma(\mathbf{x})$ and returned active-light amplitude $L_a(\mathbf{x}, \boldsymbol{\omega}_o)$ feed into the renderings at every captured modulation frequency. The frequency-dependent factors are the phasor exponent $\exp(j\,4\pi f d_s/c)$ and the sensor's demodulation contrast $\eta(f) \in (0, 1]$ from Section 2, which acts as a multiplicative gain on the rendered amplitude:

$$\hat{p}_f(\mathbf{x}, \boldsymbol{\omega}_o) \;=\; \eta(f) \int_{s_n}^{s_f} \frac{T(\mathbf{x}, \mathbf{x}_s)^2}{d_s^2}\, \sigma(\mathbf{x}_s)\, L_a(\mathbf{x}_s, \boldsymbol{\omega}_o) \exp\!\left(j\,\frac{4\pi f d_s}{c}\right) \mathrm{d}s. \tag{24}$$

The shared σ and L_a across frequencies couple the renderings at $f_1, f_2, \ldots$ into a joint constraint on the scene parameters.



(a) Recovered mean rendered depth.

(b) Phasor MAE vs. modulation frequency at $s = 0$ and $s = 3$.

Fig. 5. **Density noise remains essential across frequency counts. (a)** Even at 50 frequencies, $s=0.3$ yields fluffy reconstructions: multi-frequency angular coupling alone does not break the equivalence class ($\text{lr} = 3\times10^{-4}$, 50,000 iterations). **(b)** Phasor MAE stays low across all frequencies at both $s=0$ and $s=3$; the mild rise toward high f reflects decreasing SNR against our denoised phasor reconstructions rather than fit failure. The geometric failures in (a) are therefore variations within the equivalence class.

4.2 Does multi-frequency reduce phasor-fitting ambiguity?

In principle, a second modulation frequency constrains the spread of the phasor sum (Section 3.2). The mechanism is an angular coupling between the two rendered phasor sums that makes spread solutions more expensive. However, it does not break the equivalence class, and density noise remains necessary; multi-frequency only reduces how aggressive the noise must be.

Angular coupling. For a single density $\{w_i, d_i\}$ along a ray, each per-sample arrow's phase at frequency f_k is $\theta_i^{(k)} = (4\pi f_k/c)\, d_i$. With $r = f_2/f_1$, the two phases for the same sample are linked by

$$\theta_i^{(2)} \;=\; r\,\theta_i^{(1)}, \tag{25}$$

a single parameter-free map. Decomposing each phase into a centroid $\bar{\theta}$ plus an offset, $\theta_i^{(1)} = \bar{\theta}^{(1)} + \delta_i$, gives $\theta_i^{(2)} = r\bar{\theta}^{(1)} + r\delta_i$: viewed as a localized object on the complex plane, the phasor sum arc *rotates* (centroid shifts by $(r-1)\bar{\theta}^{(1)}$) and *stretches* (offsets scale by r). The optimizer never unwraps explicitly: $e^{jr\theta_i^{(1)}}$ is well-defined for any r and the phasor loss is computed in the complex plane.

Why the equivalence class shrinks but remains. Two phasor measurements give four real constraints per ray, while a per-ray density distribution represented at N MLP samples carries N real weights as free parameters. The cross-frequency consistency in Equation (25) makes spread solutions more expensive than at a single frequency,

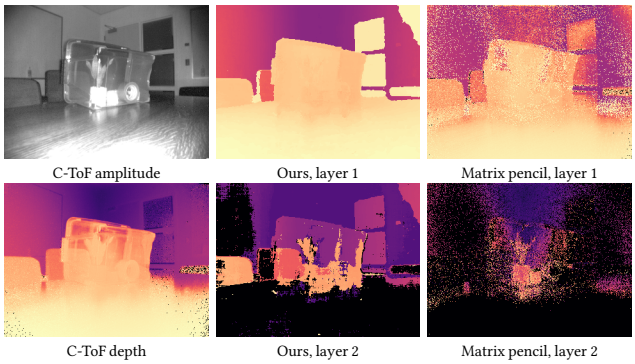


Fig. 6. **Multi-frequency separates semi-transparent layers.** Plastic box reconstruction from raw phasors at 20–70 MHz. First column: tonemapped C-ToF amplitude (top) and C-ToF depth at 20 MHz (bottom). Second and third columns: first (top) and second (bottom) depth layers from our ToF radiance field and from the pixelwise matrix-pencil baseline [11]; black indicates no valid second layer. Our scene-level optimization separates the box surface from the plant inside and yields cleaner layer estimates. Both methods produce a spurious second layer on the far wall, likely from multipath at the ceiling corner.

but for typical $N \gg 4$ the optimizer can easily find spread solutions whose weights are tuned to match both target measurements simultaneously (Figure 5). Even where the count appears favorable, such as when 50 frequencies give 100 constraints, exceeding the $N = 64$ samples typical for volume reconstructions, those constraints are coupled deterministically through Equation (25), and they bind the per-sample weights w_i, not the density σ that the optimizer controls. Since $w_i = T_i^2 \alpha_i L_a / d_i^2$ depends nonlinearly on σ through the cumulative T^2 chain and shares scale with the per-sample amplitude L_a, a spread σ can produce the same w_i as a concentrated one by trading T^2 falloff against L_a gain. Density-noise regularization again supplies the incentive that selects among these solutions.

4.3 Demonstration: Multi-layer Reconstruction

When a single pixel integrates returns from multiple depths (as happens at translucent surfaces, edges of objects, or interfaces), the spread/concentrated phasor-fitting ambiguity becomes physical: the correct density distribution *is* multi-modal. Multi-frequency angular coupling constrains the recovery of the modes' depths and weights, since a sparse mixture of returns produces a multi-frequency phasor signature distinct from any single-depth solution. We show this by estimating depth behind a translucent blue acrylic sheet from a single view (Figure 8), and by reconstructing a plant within a translucent box (Figure 6). In comparison to past works that use a per-pixel matrix pencil approach [11], our reconstruction is less noisy.

5 Multi-view Multi-frequency ToF Radiance Fields

In the multi-view multi-frequency capture this paper targets, K C-ToF cameras simultaneously observe the same scene from K different poses, each operating at its own modulation frequency f_k. The single-camera multi-frequency setting of Section 4 captured one frequency at a time. Here, all K frequencies are present in the scene at once: every camera both emits and receives, so each sensor's optical return contains not just its own active illumination at f_k but also light from every other emitter at $f_j \neq f_k$. This raises a new physical concern absent in the single-camera case: whether sensor k can still recover a clean phasor at f_k from a return that now mixes K frequencies.

Cross-frequency independence resolves the interference. The single-return model of Section 2 (Equation (1)) extends to the multi-view setting by adding leakage components from every other emitter $j \neq k$ to sensor k's returned light:

$$L_r(t) \;=\; B_0 + A\sin(2\pi f_k t - \psi) + \sum_{j\neq k} A_j \sin\big(2\pi f_j t - \psi_j\big). \tag{26}$$

Substituting into the on-chip cross-correlation (Equation (2), with reference frequency f_k) and applying the same product-to-sum reduction, we find that the desired-return term collapses to $A\cos(\psi-\phi)$ as in Equation (3), while each leakage term picks up an additional integration factor that depends on the difference frequency:

$$Q_\phi \;=\; \underbrace{A\cos(\psi-\phi)}_{\text{desired return}} \;+\; \underbrace{\sum_{j\neq k} A_j \,\mathrm{sinc}\big(\pi(f_k-f_j)\mathcal{T}\big)\cos\big(\psi_j-\phi\big)}_{\text{cross-frequency leakage}}. \tag{27}$$

Each leakage term is gated by

$$\mathrm{sinc}\big(\pi(f_k-f_j)\mathcal{T}\big), \tag{28}$$

which is unity at $f_j = f_k$ and decays with the difference frequency. With sensor frequencies separated by a few MHz and integration times $\mathcal{T} \gtrsim 1$ ms, the argument is on the order of 10^3–10^4, so $|\mathrm{sinc}(\pi(f_k - f_j)\mathcal{T})| \lesssim 3\times10^{-4}$. This sits two orders of magnitude below the OPT8241's design-target per-pixel noise floor of $\sim 4\times10^{-2}$ [34], so we can model each sensor's phasor measurement as independent of every other sensor's emission.

Multi-view multi-frequency forward model. The cross-frequency independence above is a statement about the captured measurement: $p_{f_k,k}$ at sensor k depends only on its own emitted light, even though the scene is bathed in K frequencies at once. We can therefore predict each sensor's phasor as if it were a single-emitter capture at its own f_k, and the multi-view multi-frequency forward model is the single-camera multi-frequency forward model of Equation (24) applied per sensor with its own pose and frequency:

$$\hat{p}_{f_k,k}(\mathbf{x}_k,\boldsymbol{\omega}_k) \;=\; \eta(f_k)\int_{s_n}^{s_f} \frac{T(\mathbf{x}_k,\mathbf{x}_s)^2}{d_s^2}\,\sigma(\mathbf{x}_s)\,L_a(\mathbf{x}_s,\boldsymbol{\omega}_k)\exp\!\left(j\frac{4\pi f_k d_s}{c}\right)\mathrm{ds}, \tag{29}$$

for sensor k at pose $(\mathbf{x}_k,\boldsymbol{\omega}_k)$ at modulation frequency f_k. The shared scene density $\sigma(\mathbf{x})$ and amplitude $L_a(\mathbf{x},\cdot)$ across sensors are what couple the per-sensor measurements into a joint constraint. The supervision loss extends the amplitude-normalized phasor loss of Equation (23) by summing across rays *and* sensors:

$$\mathcal{L}_{\text{multi}} \;=\; \sum_{k=1}^{K}\sum_{\text{rays of sensor } k} \mathcal{L}_{\text{phasor}}\big(\hat{p}_{f_k,k},\, p_{f_k,k}\big). \tag{30}$$

What multi-view adds to the equivalence class. On top of the angular coupling of Section 4.2, multi-view adds a complementary geometric constraint. Every 3D point $\mathbf{x}$ along a ray of sensor k is also intersected by rays from other sensors that observe the same scene volume, and the density $\sigma(\mathbf{x})$ at that point must simultaneously satisfy all those

rays' phasor measurements across varying frequencies. The optimizer fits one (σ, L_a) field against $K \times H \times W$ phasor measurements, each constrained by its own per-ray angular coupling *and* by cross-view geometric consistency. The two effects are complementary: angular coupling tightens the solution space for each individual ray (Section 4.2), and cross-view consistency tightens the solution space for the shared 3D field. Density-noise regularization (Section 3.2) is *still* necessary, but carries less of the burden.

5.1 Demonstration: Scene Unwrapping

With two C-ToF cameras, we test the multi-view multi-frequency forward model on a scene whose true depth exceeds the single-frequency unambiguous range $c/(2f)$ of any individual modulation frequency. Figure 7 shows that this approach can generate highly detailed scene reconstructions with low noise, thin features, and extended depth range even though the input raw phasors are highly noisy. Analytic depth unwrapping is unable to handle this case because each frequency is emitted from and returned to a different viewpoint.

6 Discussion

Multi-view as the real constraint, and what this enables. A theme of our analysis is that the per-ray mechanisms of density noise (Section 3.2) and multi-frequency angular coupling (Section 4.2) shrink the equivalence class but do not break it; consistent with RGB-only imaging, cross-view geometric consistency on the shared 3D field (Section 5) is where most of the depth-recovery burden actually sits. Prior C-ToF radiance-field methods [1, 5] sourced this multi-view diversity from small-baseline handheld motion of a single C-ToF camera, which aids triangulation but limits scene scale. Beyond multi-camera C-ToF depth fusion [19], wide-baseline simultaneous multi-camera multi-frequency C-ToF volume integration opens the door to detailed dynamic reconstructions because it is effectively 'single-shot' multi-frequency, e.g., for large heterogeneous sensor setups.

Reconstruction through translucent media. The front-T^2 mechanism extends to multi-layer scenes: residual transmittance past a partially transmissive surface lets the optimizer place other concentrated peaks where phasor content remains unexplained, while density noise suppresses spread between the layers and multi-frequency angular coupling (Section 4.2) penalizes this two-peak configuration relative to a centroid-equivalent spread. Therefore, the same model and loss recover opaque and translucent scenes without a separate multi-layer solver, in contrast to per-pixel spectral estimation using matrix-pencil deconvolution [11], which is neither 3D-consistent across views nor does it directly produce a shared 3D scene representation. Future work could examine how this extends to N-layer reconstruction under more realistic sensor models, and whether single-shot multi-camera captures (Section 5) could provide the required multi-frequency constraints to quickly and better image through translucent media.

Gaussian representations. Our explanation for why density noise favors concentrated solutions is specific to MLP-based volume rendering. In this setting, noise is added to the raw density output σ^{raw} before the ReLU nonlinearity (Section 3.2), which creates the front-T^2 effect described above. Although the same phasor ambiguity still exists in 3D Gaussian splatting [18, 23], the right mechanism that chooses one solution over another might be different, e.g., heuristic optimization dynamics [23], opacity-noise injection, or Gaussian dropout [7].

7 Related Work

Scene Reconstruction from RGB Images. Reconstructing 3D scenes from RGB images is a long-studied problem in computer vision and graphics. Neural radiance fields [27] set a new bar for novel-view synthesis quality, with follow-ups such as Mip-NeRF [2] and Zip-NeRF [3] addressing unbounded scenes and aliasing. 3D Gaussian Splatting [18] closed the gap on real-time rendering. Other extensions improve surface reconstruction [13, 15] or capture view-dependent and reflective effects [35, 36, 39]. RGB-only approaches rely on having many input images (typically dozens to hundreds) to recover a high-quality scene; with only a handful of views, optimization becomes severely under-constrained [7]. Sparse-view methods address this by injecting external priors: sparse depth from structure-from-motion [9], monocular depth estimates [22, 32, 42], learned two-view stereo correspondences [8], or generative regularization on unobserved views [29]. More recently, feedforward methods [6, 16, 37, 38, 41] have collapsed per-scene optimization into a single inference pass, reducing reconstruction time by orders of magnitude.

Scene Reconstruction with Active Illumination. Active illumination provides a strong geometric cue for single- or few-view reconstruction. Consumer flash can be used; for instance, Flash-Splat [40] exploits flash/no-flash cues within a Gaussian-splatting pipeline to remove unwanted reflections. Structured-light methods take a different approach: Mirdehghan et al. [28], Shandilya et al. [33] fit a neural field directly to raw structured-light images and decompose the result into direct, indirect, and ambient components. Two-bounce flash-lidar from a single view can recover otherwise-occluded geometry [20, 21]. Malik et al. [24] extend time-resolved neural radiance caching to flash-lidar captures, recovering geometry even under strong indirect illumination. SPADs push these representations into the photon-counting regime: Jungerman et al. [17] fit a NeRF directly to binary SPAD streams, and Nousias et al. [30] passively recover depth and laser localization from incidental scattering of out-of-view pulsed lasers. More directly tied to time-resolved imaging, Malik et al. [25] reconstruct radiance fields from picosecond-resolved single-photon lidar transients, matching the full pulse waveform rather than only the recovered depth.

Scene Reconstruction with C-ToF Cameras. C-ToF cameras are inexpensive, widely deployed in consumer devices (smartphones, Kinect Azure), and produce dense per-pixel depth at video frame rate. Two persistent challenges complicate this recovery: phase wrapping, where each phase is recovered only modulo 2π and caps the unambiguous depth range; and multi-path interference, where a single pixel integrates returns from several scene paths and biases the recovered depth, particularly at corners and on translucent surfaces [12]. Multi-frequency methods address both: combining wrapped phases at several modulation frequencies resolves the depth ambiguity [10], and exploiting the frequency-dependent signature of multi-path returns enables joint recovery of a sparse mixture of returns [4, 11]. For neural and related methods, TöRF [1] first optimized a NeRF directly against the phasor measurements of a C-ToF camera for dynamic

scenes. Chang et al. [5] replace the volumetric NeRF backbone with an SDF, model the C-ToF camera's active IR illumination via a learned neural lighting function, and introduce a wrapping-aware loss to resolve phase ambiguity. Later works apply to fast-moving scenes by jointly solving for scene flow and radiance from raw quad captures [31], and adapt the approach to 3D Gaussian splats for efficiency [23]. Our work analyzes the assumptions underlying this line of work and extends them to multi-frequency phasors.

References

[1] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. 2021. TöRF: Time-of-Flight Radiance Fields for Dynamic Scene View Synthesis. In *NeurIPS*, Vol. 34.

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *ICCV*. doi:10.1109/ICCV48922.2021.00580

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In *ICCV*.

[4] Ayush Bhandari, Achuta Kadambi, Refael Whyte, Christopher Barsi, Micha Feigin, Adrian Dorrington, and Ramesh Raskar. 2014. Resolving Multipath Interference in Time-of-Flight Imaging via Modulation Frequency Diversity and Sparse Regularization. *Optics Letters* 39, 6 (2014), 1705–1708.

[5] Wenjie Chang, Hanzhi Chang, Yueyi Zhang, Wenfei Yang, and Tianzhu Zhang. 2025. Learning Neural Scene Representation from iToF Imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27937–27946.

[6] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. In *CVPR*.

[7] Kangjie Chen, Yingji Zhong, Zhihao Li, Jiaqi Lin, Youyu Chen, Minghan Qin, and Haoqian Wang. 2026. Quantifying and alleviating co-adaptation in sparse-view 3d gaussian splatting. *Advances in Neural Information Processing Systems* 38 (2026), 115939–115968.

[8] Dongyoung Choi, Jaemin Cho, Woohyun Kang, Hyunho Ha, James Tompkin, and Min H. Kim. 2026. Splat-based Gradient-domain Fusion for Seamless View Transition. In *Proc. Int. Conf. 3D Vision (2026)*. Vancouver, BC, Canada.

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*.

[10] David Droeschel, Dirk Holz, and Sven Behnke. 2010. Multi-frequency Phase Unwrapping for Time-of-Flight Cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1463–1469.

[11] Micha Feigin, Ayush Bhandari, Shahram Izadi, Christoph Rhemann, Mirko Schmidt, and Ramesh Raskar. 2015. Resolving multipath interference in kinect: An inverse problem approach. *IEEE Sensors Journal* 16, 10 (2015), 3419–3427.

[12] Mohit Gupta, Shree K. Nayar, Matthias B. Hullin, and Jaime Martin. 2015. Phasor Imaging: A Generalization of Correlation-Based Time-of-Flight Imaging. *ACM Trans. Graph.* 34, 5 (2015), 156:1–18. doi:10.1145/2735702

[13] Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In *CVPR*.

[14] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. 2012. *Time-of-Flight Cameras: Principles, Methods and Applications*. Springer. doi:10.1007/978-1-4471-4658-2

[15] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH*.

[16] Roni Itkin, Noam Issachar, Yehonatan Keypur, Xingyu Chen, Anpei Chen, and Sagie Benaim. 2026. GlobalSplat: Efficient Feed-Forward 3D Gaussian Splatting via Global Scene Tokens. (2026). arXiv:2604.15284.

[17] Sacha Jungerman, Aryan Garg, and Mohit Gupta. 2026. Radiance Fields from Photons. *ACM Transactions on Graphics* 45, 1 (2026), 11:1–11:16. doi:10.1145/3770578

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139:1–14. doi:10.1145/3592433

[19] Young Min Kim, Christian Theobalt, James Diebel, Jana Kosecka, Branislav Miscusik, and Sebastian Thrun. 2009. Multi-View Image and ToF Sensor Fusion for Dense 3D Reconstruction. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1542–1549.

[20] Tzofi Klinghoffer, Siddharth Somasundaram, Xiaoyu Xiang, Yuchen Fan, Christian Richardt, Akshat Dave, Ramesh Raskar, and Rakesh Ranjan. 2025. Shoot-Bounce-3D: Single-Shot Occlusion-Aware 3D from Lidar by Decomposing Two-Bounce Light. In *SIGGRAPH Asia*. https://shoot-bounce-3d.github.io

[21] Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, and Rakesh Ranjan. 2024. PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar. In *CVPR*. doi:10.1109/CVPR52733.2024.01380

[22] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. 2024. DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization. In *CVPR*.

[23] Runfeng Li, Mikhail Okunev, Zixuan Guo, Anh Ha Duong, Christian Richardt, Matthew O'Toole, and James Tompkin. 2025. Time of the Flight of the Gaussians: Optimizing Depth Indirectly in Dynamic Radiance Fields. In *CVPR*.

[24] Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, and David Lindell. 2025. Neural Inverse Rendering from Propagating Light. In *CVPR*.

[25] Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kyros Kutulakos, and David Lindell. 2024. Transient neural radiance fields for lidar view synthesis and 3D reconstruction. In *NeurIPS*.

[26] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *CVPR*. doi:10.1109/CVPR52688.2022.01571

[27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*. doi:10.1007/978-3-030-58452-8_24

[28] Parsa Mirdehghan, Maxx Wu, Wenzheng Chen, David B. Lindell, and Kiriakos N. Kutulakos. 2024. TurboSL: Dense Accurate and Fast 3D by Neural Inverse Structured Light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 25067–25076.

[29] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *CVPR*.

[30] Sotiris Nousias, Mian Wei, Howard Xiao, Maxx Wu, Shahmeer Athar, Kevin J. Wang, Anagh Malik, David A. Barmherzig, David B. Lindell, and Kyros N. Kutulakos. 2025. Opportunistic Single-Photon Time of Flight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15852–15862.

[31] Mikhail Okunev, Marc Mapeke, Benjamin Attal, Christian Richardt, Matthew O'Toole, and James Tompkin. 2024. Flowed Time of Flight Radiance Fields. In *ECCV*.

[32] Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. 2024. CoherentGS: Sparse Novel View Synthesis with Coherent 3D Gaussians. In *Proc. ECCV*.

[33] Aarrushi Shandilya, Benjamin Attal, Christian Richardt, James Tompkin, and Matthew O'Toole. 2023. Neural Fields for Structured Lighting. In *ICCV*. 3512–3522. doi:10.1109/ICCV51070.2023.00325

[34] Texas Instruments. 2015. OPT8241 3D Time-of-Flight Sensor Datasheet. SBAS704B (Rev. B). https://www.ti.com/lit/ds/sbas704b/sbas704b.pdf

[35] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. 2022. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. In *CVPR*.

[36] Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ben Mildenhall, Benjamin Attal, Richard Szeliski, and Jonathan T. Barron. 2024. NeRF-Casting: Improved View-Dependent Appearance with Consistent Reflections. In *SIGGRAPH Asia 2024 Conference Papers*.

[37] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual Geometry Grounded Transformer. In *CVPR*.

[38] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*.

[39] Xiuchao Wu, Jiamin Xu, Chi Wang, Yifan Peng, Qixing Huang, James Tompkin, and Weiwei Xu. 2024. Local Gaussian Density Mixtures for Unstructured Lumigraph Rendering. In *SIGGRAPH*. 16:1–11. doi:10.1145/3680528.3687659

[40] Mingyang Xie, Haoming Cai, Sachin Shah, Yiran Xu, Brandon Y. Feng, Jia-Bin Huang, and Christopher Metzler. 2024. Flash-Splat: 3D Reflection Removal with Flash Cues and Gaussian Splats. In *ECCV*.

[41] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. In *ECCV*.

[42] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. 2024. FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting. In *ECCV*.

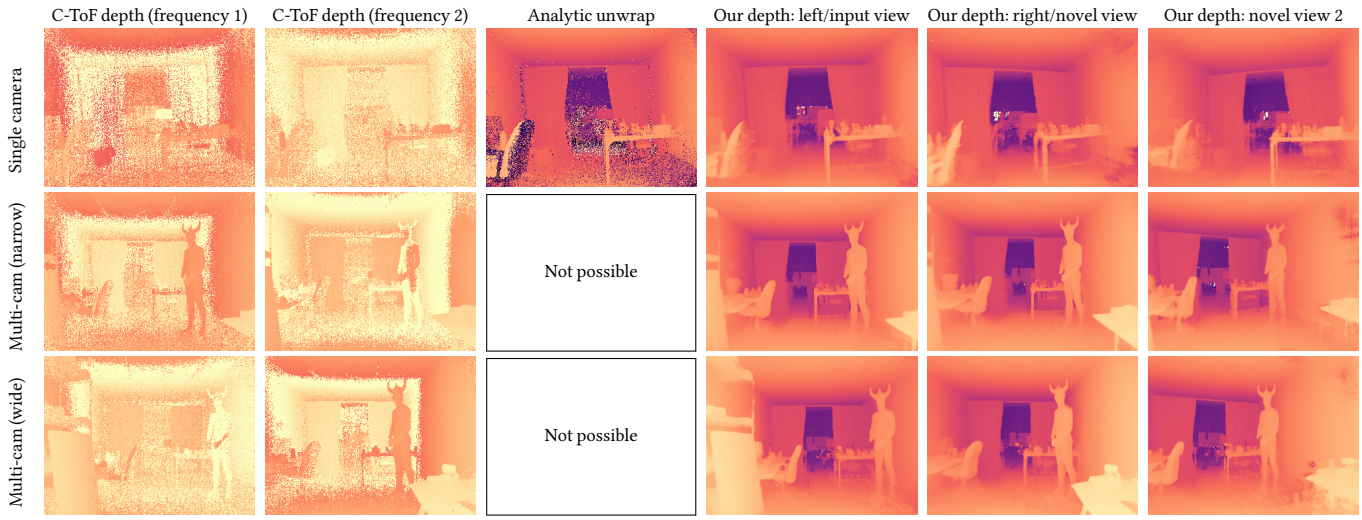


Fig. 7. **Multi-view scene unwrapping.** We fit a ToF radiance field from two C-ToF cameras modulated at 42 and 55 MHz, in three configurations (rows): a single shared camera, a narrow 10 cm baseline, and a wide 80 cm baseline. Columns 1–2 show the wrapped C-ToF depths at each input frequency; column 3 shows the analytic unwrap (only meaningful for the single-camera case); columns 4–6 show our recovered depths at the input and novel views. Despite noisy, wrapped inputs, our method recovers consistent geometry across frequencies, viewpoints, and baselines.

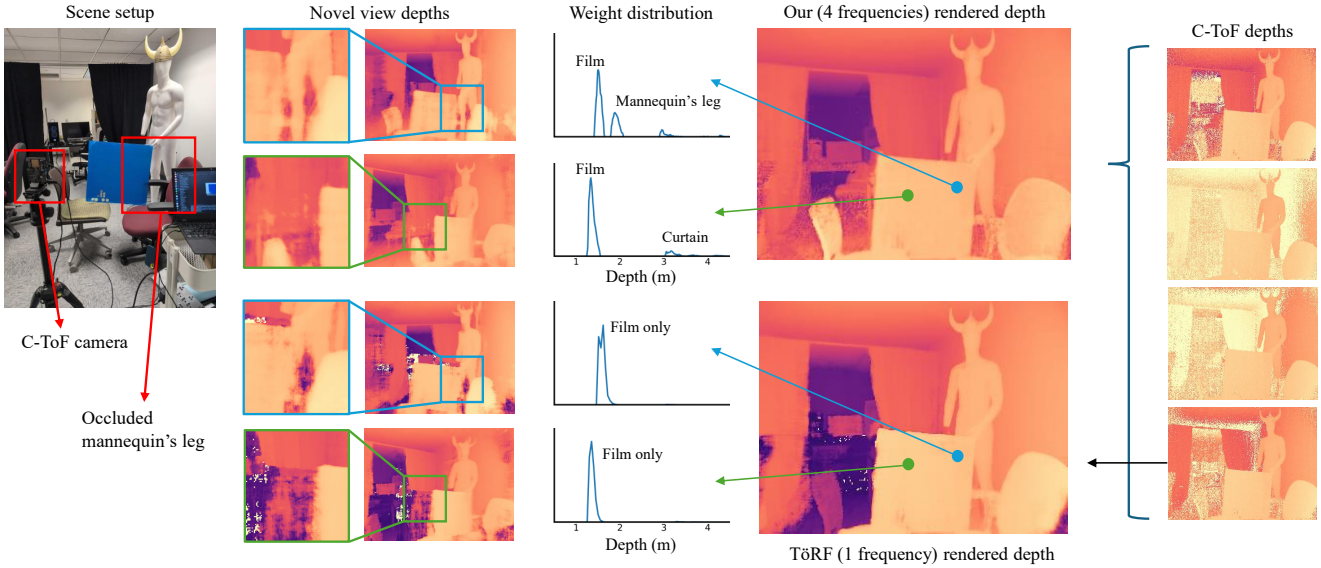


Fig. 8. **Single-view semi-transparent reconstruction from four-frequency raw C-ToF phasors.** A blue acrylic film occludes the mannequin's right leg and the background curtain in the input view. **Top:** the scene setup; our novel-view mean depths revealing the leg (upper) and curtain (lower) behind the film; per-ray volume weights at the marked pixels with distinct peaks for the film and the background; our training-view mean depth; and the raw C-ToF depths at the four input frequencies (29, 42, 55, 68 MHz). **Bottom:** TöRF [1] optimized with 29 MHz alone recovers only the foreground film and fails to separate the occluded structures.

A Phasor Fitting ≠ Mean Depth Fitting

We consider a simplified single-ray setting. Let $\theta_i = kd_i$, where $k = 4\pi f/c$, and let $w_i \geq 0$ be the contribution of sample i. The rendered phasor is

$$\hat{p} = \sum_i w_i e^{j\theta_i}, \tag{31}$$

while the rendered mean phase is

$$\bar{\theta} = \frac{\sum_i w_i \theta_i}{\sum_i w_i}. \tag{32}$$

Let $\beta_i = w_i / \sum_j w_j$, and write

$$\theta_i = \bar{\theta} + \delta_i, \qquad \sum_i \beta_i \delta_i = 0. \tag{33}$$

Then

$$\sum_i \beta_i e^{j\theta_i} = e^{j\bar{\theta}} \sum_i \beta_i e^{j\delta_i}. \tag{34}$$

If all contributing samples lie at the same depth, then $\delta_i = 0$ for all i, and the accumulated phasor is exactly $e^{j\bar{\theta}}$.

However, for a spread-out distribution, the phase of the accumulated phasor is not necessarily $\bar{\theta}$. The mean-depth condition gives $\sum_i \beta_i \delta_i = 0$, but the phasor phase depends on $\sum_i \beta_i \sin \delta_i$. Since

$$\sum_i \beta_i \sin \delta_i = -\frac{1}{6} \sum_i \beta_i \delta_i^3 + O(\delta_i^5), \tag{35}$$

this term is not guaranteed to vanish. Therefore,

$$\hat{p} = A_t e^{j\theta_t} \not\Rightarrow \bar{\theta} = \theta_t. \tag{36}$$

Thus, single-frequency phasor fitting can match the raw C-ToF measurement while still producing an incorrect rendered mean depth. This motivates optimization biases or regularization that encourages a sparse, concentrated density distribution along each ray: when the density collapses near a single depth, phasor fitting becomes better aligned with mean-depth fitting.

B Derivation of the Amplitude Weighted L2 Loss

In this section, we derive Equation (22). Let the predicted phasor be

$$\hat{\mathbf{p}} = (\hat{x}, \hat{y}), \qquad \hat{r} = \sqrt{\hat{x}^2 + \hat{y}^2}, \qquad \hat{\phi} = \operatorname{atan2}(\hat{y}, \hat{x}), \tag{37}$$

and let the captured phasor be

$$\mathbf{p} = (x, y), \qquad r = \sqrt{x^2 + y^2}, \qquad \phi = \operatorname{atan2}(y, x). \tag{38}$$

We define

$$\Delta x = \hat{x} - x, \qquad \Delta y = \hat{y} - y. \tag{39}$$

We consider the error in log-polar coordinates:

$$\Delta \log r = \log \hat{r} - \log r, \qquad \Delta\phi = \hat{\phi} - \phi. \tag{40}$$

First-order expansion of $\Delta \log r$. Define

$$g(x, y) = \log \sqrt{x^2 + y^2}. \tag{41}$$

Using a first-order Taylor expansion of $g(x, y)$ around the prediction $(\hat{x}, \hat{y})$,

$$g(x, y) \approx g(\hat{x}, \hat{y}) + \frac{\partial g}{\partial x}\Big|_{(\hat{x},\hat{y})} (x - \hat{x}) + \frac{\partial g}{\partial y}\Big|_{(\hat{x},\hat{y})} (y - \hat{y}). \tag{42}$$

Rearranging gives

$$\log \hat{r} - \log r \approx \frac{\partial g}{\partial x}\Big|_{(\hat{x},\hat{y})} (\hat{x} - x) + \frac{\partial g}{\partial y}\Big|_{(\hat{x},\hat{y})} (\hat{y} - y). \tag{43}$$

Since

$$\frac{\partial g}{\partial x} = \frac{x}{x^2 + y^2}, \qquad \frac{\partial g}{\partial y} = \frac{y}{x^2 + y^2}, \tag{44}$$

we obtain

$$\Delta \log r \approx \frac{\hat{x}\Delta x + \hat{y}\Delta y}{\hat{x}^2 + \hat{y}^2}. \tag{45}$$

First-order expansion of $\Delta\phi$. Define

$$h(x, y) = \operatorname{atan2}(y, x). \tag{46}$$

Again using a first-order Taylor expansion of $h(x, y)$ around $(\hat{x}, \hat{y})$,

$$h(x, y) \approx h(\hat{x}, \hat{y}) + \frac{\partial h}{\partial x}\Big|_{(\hat{x},\hat{y})} (x - \hat{x}) + \frac{\partial h}{\partial y}\Big|_{(\hat{x},\hat{y})} (y - \hat{y}), \tag{47}$$

which gives

$$\hat{\phi} - \phi \approx \frac{\partial h}{\partial x}\Big|_{(\hat{x},\hat{y})} (\hat{x} - x) + \frac{\partial h}{\partial y}\Big|_{(\hat{x},\hat{y})} (\hat{y} - y). \tag{48}$$

Since

$$\frac{\partial h}{\partial x} = -\frac{y}{x^2 + y^2}, \qquad \frac{\partial h}{\partial y} = \frac{x}{x^2 + y^2}, \tag{49}$$

we obtain

$$\Delta\phi \approx \frac{-\hat{y}\Delta x + \hat{x}\Delta y}{\hat{x}^2 + \hat{y}^2}. \tag{50}$$

Combining the two terms. Substituting Equation (45) and Equation (50) gives

$$(\Delta \log r)^2 + (\Delta\phi)^2 \approx \frac{\left(\hat{x}\Delta x + \hat{y}\Delta y\right)^2}{(\hat{x}^2 + \hat{y}^2)^2} + \frac{\left(-\hat{y}\Delta x + \hat{x}\Delta y\right)^2}{(\hat{x}^2 + \hat{y}^2)^2}. \tag{51}$$

Using the identity

$$(au + bv)^2 + (-bu + av)^2 = (a^2 + b^2)(u^2 + v^2), \tag{52}$$

with

$$a = \hat{x}, \qquad b = \hat{y}, \qquad u = \Delta x, \qquad v = \Delta y, \tag{53}$$

we obtain

$$(\Delta \log r)^2 + (\Delta\phi)^2 \approx \frac{\Delta x^2 + \Delta y^2}{\hat{x}^2 + \hat{y}^2} = \frac{\Delta x^2 + \Delta y^2}{\hat{r}^2}. \tag{54}$$

This is the local approximation used to motivate Equation (23).

C Capture Details

Cameras and Calibration. We capture with a pair of Texas Instruments OPT8241-EVM C-ToF cameras. Per-camera intrinsics are calibrated on a planar checkerboard. Multi-camera extrinsics are recovered from retroreflective spheres that saturate the raw quads, giving high-contrast point correspondences across views; we estimate essential and fundamental matrices from these correspondences and then jointly refine intrinsics and extrinsics by bundle adjustment. The per-pixel phase offset is calibrated against a planar surface at known depth, and fixed-pattern noise is estimated as a per-quad mean over dark frames and subtracted. We calibrate the sensor's demodulation contrast $\eta(f)$ as a single global scalar per modulation frequency from a white-wall frequency sweep, expressed relative to the 40 MHz channel; we do not calibrate higher-order non-idealities

such as amplitude non-linearity at low frequencies or per-pixel contrast variation. For each scene, we record an ambient frame, then the calibration capture, then the scene.

Datasets. Our captured datasets consist of a single camera and multi-camera setups. For single camera scenes we perform a frequency sweep, where the camera captures 100 frames at each modulation frequency from 20 MHz to 80 MHz. For multi-camera scenes we consider a stereo pair, with both cameras placed a baseline apart facing the scene. We capture scenes both at close to medium range for single camera setups, and at medium to long range for multi-camera setups where we anticipate additional supervision from multi-view constraints to be helpful.