FirstCross: Convex Single-Photon 3D Reconstruction from Sparse Views

Chi-Wei Lee Cheng-Han Chiang Yi-Chun Chang An-Kai Chen Po-Chih Kuo
National Tsing Hua University, Hsinchu, Taiwan
arthur031221.github.io/firstcross

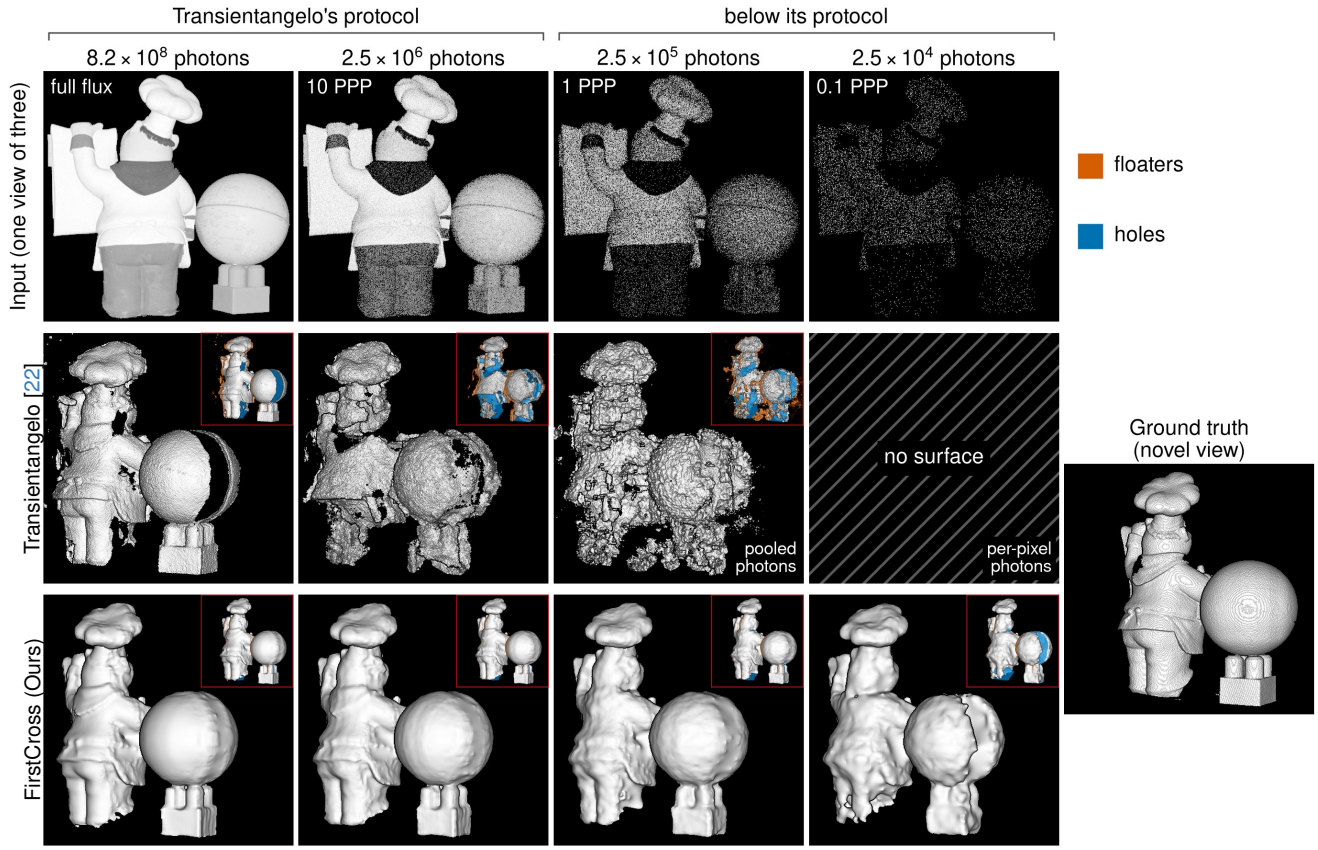


Figure 1. **FirstCross reconstructs 3D shape from very few photons.** Novel view of the real test scene chef from three views, with their total photon count above each column. PPP is the mean photon count of an occupied pixel, and 1 and 0.1 PPP are thinned from 10 PPP. Red-framed insets mark depth errors over 2 cm, as floaters (in front of the true surface or off the object) or holes (behind it or missing). Transientangelo leaves floaters and holes even at full flux. As it recovers no surface from per-pixel photons below its protocol's lowest level, 10 PPP, we show it at 1 PPP with adapted settings (Sec. S3) on photons pooled over 4×4 pixels.

Abstract

Single-photon lidar records the arrival time of each returning photon, enabling 3D reconstruction from few views and scarce photons. However, leading neural methods such as Transientangelo require over an hour of non-convex per-scene optimization, and their accuracy degrades as photon flux decreases. We propose FirstCross, a network-free approach that instead solves a short sequence of convex problems on a voxel grid. For a solid object, the Poisson likelihood of a ray with given signal and background rates depends on the scene solely through its first crossing, the point where the ray first enters the object. Decomposing this likelihood yields two tractable terms. The first penalizes surfaces in front of detected photons and admits a tight convex closure, namely volume rendering with a

Emails: {levi74108520963, faidavid7, changg587, kylechen.ankai}@gmail.com, kuopc@cs.nthu.edu.tw

running-maximum transmittance in place of NeRF's product transmittance. The second accounts for photons that no surface explains and admits a convex upper bound once the object is assumed to be thick behind each detected surface point. Compared with Transientangelo under its own real-data protocol, FirstCross achieves lower depth error across every test scene, view count, and flux level. With three views and 100 times fewer photons than at the protocol's lowest flux, our mean object depth error stays under half of Transientangelo's with all photons. Given our detected surface points, the best classical fusion has 1.3 to 1.8 times our three-view error across flux levels. On real room-scale captures with 1% of their photons, TransientNeRF has 2.8 to 7.7 times our error. FirstCross lets few-view single-photon scanners cut exposure to 1% and reconstruct an object in about a minute on one GPU.

1. Introduction

In single-photon lidar, a pulsed laser and a single-photon avalanche diode (SPAD) record a transient histogram of photon arrival times at each pixel. Each signal photon measures range, and a few of them locate a surface to within the width of the instrument response function (IRF), the timing jitter of laser and detector. However, fast acquisition, long range or low-power sensors yield fewer than one photon per pixel [11, 38], as at 0.1 photons per occupied pixel (PPP) in Fig. 1.

Prior work follows two main routes. The classical route detects a return on each ray (the arrival time of the echo from the surface) and fuses the returns [8] and the free space in front of them [4, 49] across views. A ray without a return contributes nothing, although the absence of photons is evidence of free space. The strongest neural methods, transient neural radiance fields [23, 24] and their surface variant Transientangelo [22], instead volume-render histograms from a neural field and fit the measurements bin by bin. They are accurate with many photons, but fitting each scene takes over an hour and the objective is non-convex.

We propose FirstCross, which starts instead from the photon likelihood (Fig. 2). Our key observation is that, for a solid object, photon arrivals along a ray with given signal and background rates depend on the scene only through one point: the first intersection of the ray with the object, which we call the *first crossing*. We write the Poisson likelihood of every ray, with or without photons, as a function of its first crossing and then of the occupancies of its samples in an implicit field on a voxel grid. This likelihood is combinatorial in the occupancies, but it splits into two parts. The *too-early* part, where a surface blocks the path of a photon, is submodular in the set of occupied samples. Its tight convex closure, the largest convex function that matches it on binary occupancies, has a closed form: volume rendering whose transmittance is one minus the running maximum of occupancy along the ray, in place of NeRF's product transmittance. The *too-late* part, where no surface explains a photon, is supermodular and the only source of non-convexity. It requires the ray to enter the object before the photon arrives without specifying where. Our *thickness assumption*, that objects are thicker than a few IRF widths, supplies the location, because the point just behind a detected return then lies inside the object. Each solver iteration therefore works on a convex problem, and the minimum of the first problem does not depend on the initial grid, although our solver reaches it only approximately. With the product transmittance, by contrast, an optimisation started from an occupied grid can leave free space filled. We make three contributions:

- We show that the first-crossing Poisson likelihood splits into a submodular too-early part and a supermodular too-late part. We derive the tight convex closure of the too-early part in closed form, which is volume rendering with a running-maximum transmittance.
- We derive a convex upper bound on the too-late part from the thickness assumption. Our solver needs no neural network and reconstructs an object in about a minute.
- We demonstrate lower depth error than Transientangelo on every test scene, view count and flux level of its own real-capture protocol. From three views, our error on the object stays below Transientangelo's at full flux even with inputs thinned 100-fold below its lowest flux of 10 PPP.

Scope. We assume solid objects, known camera poses and negligible pile-up, which biases detections toward early time bins at high flux. Thin sheets, partially transmissive layers [45] and mixed pixels at silhouettes are not modelled.

2. Related Work

Neural transient fields. TransientNeRF [23] introduced volume rendering of multi-view transient histograms. Transientangelo [22] then targeted sparse views and low flux with a neural signed distance field [17] and regularisers. Other work models propagating light [24, 25], outdoor scenes [37], low-cost sensors [29] and mixed returns [45], or learns to reconstruct from a single transient [16], and SP-TransientBench [5] provides room-scale captures.

Photon-efficient depth imaging. First-photon imaging [11] and later work [6, 19, 32, 34, 38, 40] use Poisson statistics for single-view depth, with spatial priors over neighbouring pixels. With several views, all first crossings lie on one surface, so our field couples rays across views.

Range fusion and implicit surfaces. Volumetric fusion [4] and its robust convex forms [49] combine per-ray distances. Poisson-type reconstruction [2, 8, 9, 33] and radial basis functions [3] fit an implicit function to oriented points, and variants estimate orientation [7, 18] or add visibility envelopes [10]. Our bending prior follows [2, 3], but our data term lets zero-photon rays carve free space and

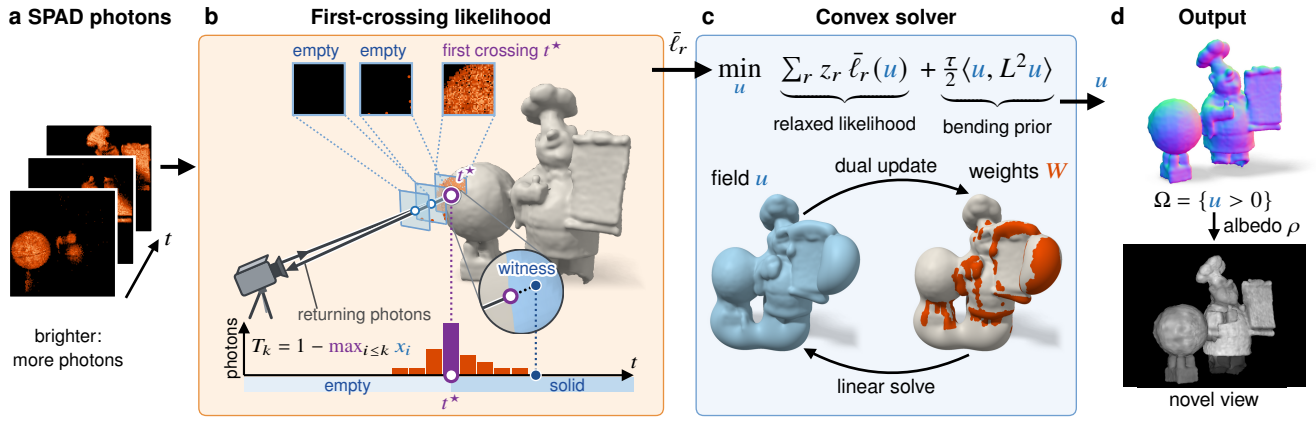


Figure 2. **Overview of FirstCross on real data.** FirstCross fits a field u by a short sequence of convex problems. (a) Photons of one view at 1 PPP in three arrival-time slices. (b) The likelihood of ray r depends on the scene only through its first crossing $t^\star$ (purple). Cross-sections across r (top) show almost no photons in free space, and the pooled photons of r (bottom) peak at $t^\star$. The witness (dark blue), just behind the return, is assumed to lie inside the object. (c) Each of the 12 iterations puts the weights W (orange) where rays enter the current object, then updates u by a linear solve. The weight z_r of ray r falls with its photon count. (d) The shape $\Omega = \{u > 0\}$ and a novel view.

weights rays by effective photon counts.

Visibility in surface reconstruction. Space carving [13] and continuous multi-view stereo [12] use per-voxel photo-consistency, and s–t cuts [14] and VASE [41] use soft free-space constraints from points. Closest to ours are ray potentials [20, 35, 42], whose cost depends on a ray's first occupied voxel and is minimised discretely or by a convex relaxation with a non-convex visibility constraint [36]. Concurrently, Lenain *et al.* [15] regress the first intersection of a 1-Lipschitz network, a non-convex problem. We share this first-crossing structure and submodular tools [1, 21, 30], but our cost is the photon likelihood, and we derive the tight convex closure of its too-early part and a convex bound on its too-late part. This closure is a running-maximum rendering that models visibility in a convex functional.

Volume rendering of surfaces. NeRF [26] composites independent opacities. NeuS [44], VolSDF [47] and UNISURF [31] derive them from an implicit surface but keep the product along the ray. NeuS matches the running maximum on rays that cross one surface. IDR [46] renders the first intersection alone, and non-exponential transmittance [39, 43] and stochastic geometry [27] model opaque surfaces. All of these agree on binary occupancy and differ on fractional values. The running maximum is convex, and Proposition 1 shows that no convex extension is tighter.

3. From Photons to a Convex Functional

From a few transient histograms with known poses, we seek a field u on a voxel grid that is positive exactly inside the object. We write the likelihood of a ray through its first crossing, replace its too-early and too-late parts with convex surrogates, and minimise the resulting functional (Fig. 2). Sec. S1 gives proofs and derivations.

3.1. First-Crossing Likelihood

Let $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ be a camera ray, with time measured as one-way range, and let its first crossing $t^\star$ be the first depth where $u > 0$, or infinity if there is none. Neglecting pile-up, photon arrivals on the ray form a Poisson process with intensity $a\,h(t-t^\star)+b$, where the IRF h is a Gaussian of standard deviation σ, a is the expected signal count and b is the background rate. Keeping only the larger of $a\,h$ and b in the log-intensity of each photon changes its cost by at most $\log 2$. Up to constants, the negative log-likelihood of the arrival times $t_1, \dots, t_n$ then becomes

$$\ell(t^\star) = a\,\mathbf{1}[t^\star < \infty] + \sum_{j=1}^{n} \min\Big\{\frac{(t_j - t^\star)^2}{2\sigma^2},\ \omega\Big\}, \qquad (1)$$

where $\omega = \log(a\,h(0)/b)$, the cost in nats of an unexplained photon, is held constant. A crossing before the arrival of a photon blocks its path, a later or missing crossing leaves the photon unexplained, and a crossing near the photon incurs a quadratic cost. A ray with no photons pays its expected signal count for any crossing and thus carves free space.

3.2. Convex Closure of the Too-Early Cost

To optimise this cost, we sample the ray at depths $\xi_1 < \cdots < \xi_K$ and let the binary vector x mark the samples inside the object. The cost depends on x only through the first occupied sample. It is thus a set function $F(A) = \ell(\min A)$ of the set A of occupied samples, where $\ell(k)$ is the cost of a crossing at sample k, and $K+1$ means no crossing. If ℓ is non-increasing on $\{1, \dots, K+1\}$, then F is submodular, $F(A) + F(B) \ge F(A \cup B) + F(A \cap B)$, and if ℓ is non-decreasing, then F is supermodular.

Every such cost is the sum of a non-increasing and a non-decreasing part, and we split Eq. (1) one photon at a time,

$$\ell^{\downarrow}(t) = a\,\mathbf{1}[t < \infty] + \textstyle\sum_j \mathbf{1}[t < t_j]\,\phi_j(t),$$
$$\ell^{\uparrow}(t) = \textstyle\sum_j \mathbf{1}[t > t_j]\,\phi_j(t). \tag{2}$$

Here ϕ_j is the cost of photon j in Eq. (1). The too-early cost $\ell^{\downarrow}$ penalises a surface on the ray before t_j, whereas the too-late cost $\ell^{\uparrow}$ penalises a missing surface by t_j. By monotonicity, $\ell^{\downarrow}$ makes F submodular and $\ell^{\uparrow}$ supermodular.

Proposition 1 (Tight convex closure)**.** *Let $\ell^{\downarrow}$ be non-increasing with $\ell^{\downarrow}(K+1) = 0$, $F(A) = \ell^{\downarrow}(\min A)$, $\Delta_k = \ell^{\downarrow}(k) - \ell^{\downarrow}(k+1) \geq 0$ and $\bar{x}_k = \max_{i \leq k} x_i$, the running maximum of x. Then the Lovász extension of F is*

$$F^L(x) = \sum_{k=1}^{K} \Delta_k\, \bar{x}_k, \qquad x \in [0,1]^K. \tag{3}$$

It is convex, equals F on binary x, and is the largest convex function on $[0,1]^K$ with these values.

The closure follows from the Lovász extension formula [1, 21], which averages the set function over the level sets of the occupancies. A level set starts at or before a sample exactly when the running maximum at that sample reaches the level, and the closure is the expected too-early cost of the first crossing of a random level set. It stays convex when the occupancies are a convex, non-decreasing function g of the field. As clipping g at one would break convexity, occupancies may exceed one inside the object. This costs nothing on rays with a return, where we drop the expected signal count, and the too-early cost stops decreasing behind the last photon (Sec. S1). The closure is tight for one ray and, summed over rays, a convex surrogate in u.

Running-maximum transmittance. For $x \in [0,1]^K$, let $T_k = 1 - \bar{x}_k$, $T_0 = 1$ and $\pi_k = T_{k-1} - T_k$. Then $F^L(x) = \sum_k \pi_k\, \ell^{\downarrow}(k)$, which is volume rendering with transmittance one minus the running maximum. NeRF's transmittance $T_k = \prod_{i \leq k}(1 - x_i)$ and ours are two extensions of one binary function, where the opacity $1 - T_k$ is the OR of the samples up to k. The product is this OR for independent samples, as in a participating medium, and the maximum for samples thresholded at one shared level, as along a ray through a solid object.

With either transmittance, the too-early cost is a non-negative combination of the opacities, but the two differ in the gradient on a floater (Fig. 3b, c). Under the product, the derivative of an opacity with respect to the occupancy of one sample is the transmittance of the other samples, which nearly vanishes for a dense floater several samples thick, however many photons it blocks. Under the running maximum, the cost is convex with a non-decaying subgradient, and the floater is removed at a rate proportional to the photons it blocks. This argument does not cover per-bin losses, which act on a floater through the histogram it renders.

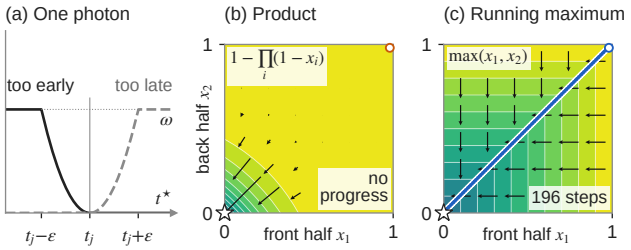


Figure 3. **The too-early cost and its two extensions.** (a) The cost of a photon at t_j against the first crossing $t^\star$ splits into too-early and too-late costs, capped at ω beyond $t_j \pm \varepsilon$. (b, c) The too-early cost (darker is lower) of an eight-sample floater over the occupancies x_1, x_2 of its halves, both empty at the star. Gradient descent from a nearly opaque floater (dot) stalls for 20 000 steps under the product and empties it in 196 under the running maximum.

3.3. Convex Bound on the Too-Late Cost

The too-late cost is a non-negative combination of the concave functions $1 - \bar{x}_k$ by summation by parts. A photon only indicates that some sample near its arrival lies inside the object. Any such sample i is a witness, since $1 - \bar{x}_k \leq [1 - x_i]_+$ for $i \leq k$, where $[\cdot]_+$ is the positive part. Each choice of witnesses thus bounds the concave term by a convex one. In place of the tangent of the concave–convex procedure [30, 48], we use a convex bound in the field, which does not guarantee monotone descent.

In the first iteration, the field does not yet show where the object is, and the thickness assumption provides the witness. A photon can be explained within the band $\varepsilon = \sqrt{2\omega}\,\sigma$ around it, at whose edge its cost reaches ω (Fig. 3a). If the object is thicker than this band along the ray, the point one band behind the detected return $\hat{t}_r$ (Sec. 3.6) is inside it. In later iterations, we use the current field to place the witness between half a band and one band behind the return, at the distance s_r that minimises Eq. (4). For a ray with N_r photons in its return and the witness $\boldsymbol{p}_r^+ = \boldsymbol{r}(\hat{t}_r + s_r)$, the unexplained-photon cost ωN_r becomes

$$B_r(u) = \frac{\omega N_r}{m_r^2} \left(\frac{\varepsilon}{s_r}\right)^2 \left[\, \frac{m_r s_r}{\varepsilon} - u(\boldsymbol{p}_r^+) \right]_+^2. \tag{4}$$

Here $m_r > 0$ is the scale of u that we define in Sec. 3.5. The term is at least ωN_r while the witness lies outside the object and vanishes once u reaches $m_r s_r / \varepsilon$ there. It is thus a convex upper bound on that cost.

3.4. Surface Localisation

To localise the surface within the band, we take the occupancy near a photon at t_j to be the ramp $x(t) = [(t - t_0 + w)/(2w)]_+$ of half-width w about t_0, with offset $\delta = t_0 - t_j$ and $|\delta| < w$, and ignore the cap ω. The continuous too-early and too-late costs of this photon are then $(w - \delta)^3/(12 w \sigma^2)$ and $(w + \delta)^3/(12 w \sigma^2)$, whose

sum is

$$\frac{\delta^2}{2\sigma^2} + \frac{w^2}{6\sigma^2}. \tag{5}$$

For every ramp width, the sum is the cost of a sharp crossing at t_0 plus a term that does not depend on the offset, which has two consequences. First, the slope $\partial_t u$ of u along the ray sets the ramp width but not the location accuracy, and no eikonal term is required. Second, for a crossing at $t^\star$, the N_r photons of a return have, up to a constant, the cost of N_r photons at their centroid $\hat{t}_r$. Linearising u along the ray at $\boldsymbol{p}_r = \boldsymbol{r}(\hat{t}_r)$ gives the Gauss–Newton form that we discretise,

$$C_r(u) = \frac{N_r\, u(\boldsymbol{p}_r)^2}{2(\partial_t u)^2 \sigma^2}. \tag{6}$$

3.5. Overall Functional

With a per-ray weight z_r defined below, the functional is

$$\begin{aligned} E(u) &= \sum\nolimits_r z_r\, \bar{\ell}_r(u) + \tfrac{\tau}{2}\langle u, L^2 u\rangle, \\ \bar{\ell}_r(u) &= \sum\nolimits_k \Delta_{r,k}\, \bar{x}_{r,k} + B_r + C_r, \\ x_{r,k} &= g\big((u(\boldsymbol{r}(\xi_k)) + m_r)/m_r\big), \\ \Delta_{r,k} &= \ell_r^{\downarrow}(\xi_k) - \ell_r^{\downarrow}(\xi_{k+1}), \\ \ell_r^{\downarrow}(t) &= \omega\,[n_r(t) - q\, n_r]_+ + a_r(t), \end{aligned} \tag{7}$$

where $g(y) = 2(\sqrt{1 + [y]_+^2} - 1)$ is convex and non-decreasing, as the closure requires. The *too-early term* carries the capped part $\ell_r^{\downarrow}$ of the too-early cost. Here $n_r(t)$ counts the photons after $t + \varepsilon$ among all $n_r \geq N_r$ photons of the ray, q is the fraction of photons allowed behind the crossing for multipath, mixed pixels and the IRF tail, and $a_r(t)$ is zero on rays with a return and otherwise the expected signal count of a crossing at t, which falls off with the square of t (Sec. S2). The *too-late term* B_r carries the capped part of the too-late cost, and the *location term* C_r carries the quadratic parts of both costs in the band around each return photon, summed in closed form by Eq. (5). Outside the band the location term stays quadratic, as a cap would break convexity. The bending prior, with strength τ and discrete Laplacian L, fills in unobserved regions. All terms are convex in u for fixed slopes, scales and witnesses.

Scale of the field. The likelihood depends only on the zero level set of u. The too-early and too-late terms fix the slope of u along the ray at m_r/ε (Sec. S1). We set $m_r = \beta\varepsilon c_r$ with the incidence cosine c_r of the ray, so that a plane of slope β along its normal has slope m_r/ε along every ray.

Effective count. Our likelihood treats the photons of a ray as independent draws from an exact model. Errors of calibration, pulse shape, residual pile-up and mixed pixels do not shrink with the photon count and can outweigh photon noise at full flux (Sec. S5). We therefore coarsen the likelihood of each ray [28], which approximately raises it to the power $z_r = \alpha/(\alpha + n_r)$ (Sec. S1). The effective count $z_r n_r$ stays below α, and the functional stays convex.

Parameters. Measuring lengths in units of σ makes the functional dimensionless, and the band ε is fixed by ω. From objects with $\sigma = 4.4$ mm to rooms with $\sigma = 9$ cm, we use $\omega = 3.7$ nats, $\beta = 0.5$, $\tau = 2.646$, $\alpha = 20$ and $q = 0.1$. The grid has 256^3 voxels for objects and 384^3 for rooms.

3.6. Return Detection

The statistics in the functional are computed from the histograms with standard detection steps (Sec. S2). On objects, each ray pools the photons of 4×4 pixels (one pixel on rooms), while depth is still rendered at every pixel. We take the return $\hat{t}_r$ of a ray as the centroid of its earliest strong matched-filter response, and a ray has a return if its local false-discovery rate against simulated background-only rays is at most $1/2$, with no threshold tuned per scene or flux. Rays without a return are treated as zero-photon rays. The full-flux captures are provided with object masks applied, and at low flux we use the masks of Transientangelo, giving both methods identical inputs.

3.7. Solver and Rendering

To minimise E, we alternate a dual update along rays with a linear solve for u. The dual update writes each running maximum as a maximum over distributions of where the ray enters the object and gives each sample a weight W. With these weights fixed, we replace each term by a quadratic model (Alg. S1). We run 12 iterations from the empty grid $u = -\beta\varepsilon$ with unit slope and cosine. Each iteration is one damped dual update and one solve by FFT-preconditioned conjugate gradients, after which the slopes, cosines, scales and witnesses are updated. Keeping all four at their first-iteration values gives the single convex problem of Tab. 2.

Depth and intensity rendering. Depth is the first zero crossing of the field along the ray, refined by linear interpolation, or the location of its maximum if there is no zero crossing. Novel views and transients are rendered from the first crossing with a per-point maximum-likelihood albedo ρ (Sec. S2) and without a learned appearance model.

4. Experiments

Datasets and protocol. We follow the protocol of Transientangelo [22]: its real captures, two, three or five input views, ten novel views, full flux and its low-flux settings of 300 to 10 PPP. Transientangelo results are obtained from its released checkpoints, which reproduce 14 of its 16 published three-view results (Sec. S3). We further simulate 1 and 0.1 PPP, and down to 0.003 PPP in Fig. 4, by binomial thinning of 10 PPP, which keeps its signal-to-background ratio. The official depth error averages $|\hat{d} - d|$ over the frame, with zero outside the ground-truth mask. On baskets and food, this mask includes the turntable's back wall. We

Table 1. **Depth error (mm, ↓) on the test scenes of the Transientangelo (TA) protocol.** Ours is lower than TA in every pair. TA uses its released checkpoints, none below 10 PPP, and 1 and 0.1 PPP are thinned from 10 PPP. Rows use three views unless stated, and the object depth MAE excludes the back wall. †The boots checkpoint misses most of the object here in at least half of the novel views, and ours is also lower without boots. The TransientNeRF row averages the per-scene errors rounded to 1 mm in [22], with – where none is reported.

Flux (PPP)	Full		300		150		50		10		1	0.1
Scene	TA [22]	Ours	TA	Ours	TA	Ours	TA	Ours	TA	Ours	Ours	Ours
baskets	7.58	**5.73**	7.26	**5.63**	7.23	**5.68**	7.55	**5.60**	14.77	**5.66**	5.67	5.64
boots	0.92	**0.42**	0.86	**0.41**	0.83	**0.41**	33.21†	**0.41**	0.62	**0.42**	0.43	0.58
chef	3.88	**0.91**	3.50	**0.95**	4.81	**0.92**	2.73	**0.94**	5.17	**0.94**	1.08	1.76
food	6.43	**4.84**	6.30	**4.84**	6.26	**4.85**	5.94	**4.86**	7.62	**4.86**	4.94	5.02
Mean	4.70	**2.97**	4.48	**2.96**	4.78	**2.97**	12.36†	**2.95**	7.04	**2.97**	3.03	3.25
TransientNeRF [23] from [22]	5.75		6.50		17.50		–		48.00		–	–
Mean, 2 views	6.34	**3.07**	3.92	**3.11**	5.09	**3.06**	16.47†	**3.09**	7.67	**3.10**	3.28	4.02
Mean, 5 views	6.02	**2.79**	13.84†	**2.80**	5.36	**2.80**	7.99	**2.81**	6.54	**2.81**	2.82	2.95
Mean object depth MAE	2.60	**0.90**	2.36	**0.88**	2.67	**0.89**	10.23†	**0.88**	4.94	**0.89**	0.95	1.17

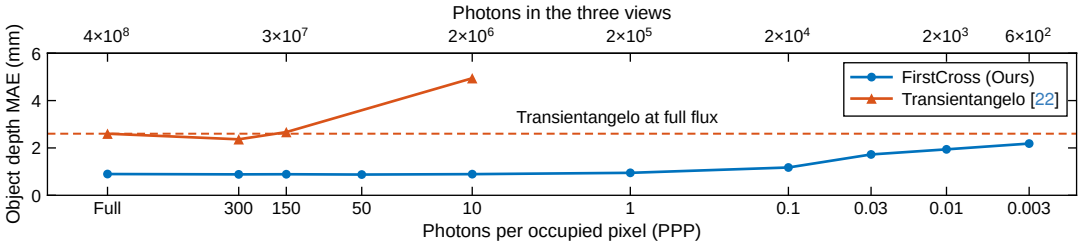


Figure 4. **Mean object depth MAE against photon count over the three-view test scenes.** Ours stays below Transientangelo's full-flux error (dashed) down to fewer than 10^3 photons in total. Transientangelo at 50 PPP is omitted, as its boots checkpoint fails.

therefore also report the object depth MAE, which further zeroes the error beyond 0.30 m from the scene origin. For room scale, we use the real rooms of SP-TransientBench [5] at full flux and thinned to 1%. All parameters are set on the validation scenes cinema, carving and library_floor2. We report single runs, and Sec. S5 adds repeated runs, other thinnings, extra ambient photons and camera rotations.

Comparison to Transientangelo. Table 1 and Tab. S4 show that ours has the lower official depth error on every test scene and flux level of the protocol with two, three and five views. With three views, our mean error is 2.97 mm at both full flux and 10 PPP, whereas Transientangelo's rises from 4.70 to 7.04 mm. This is a reduction of 37% and 58%, respectively, and Sec. S2 gives bootstrap intervals. Our return detection accounts for much of the gap: TV-L1 fusion [49] of our detected returns has less than half the object depth MAE of Transientangelo at full flux and 10 PPP (Tabs. 1 and 2). However, the best classical fusion has 1.3 to 1.8 times our mean object depth MAE at every flux level in Tab. S6. On an RTX 4090, ours takes 51 s in total on chef at 10 PPP, and Transientangelo trains for 1 h 26 min.

In Fig. 5 and Tab. S8, the two methods differ most in holes and floaters, which dominate the mean error, but their medians are close at full flux. At 10 PPP, Transientangelo's median grows 2.7 and 5.4 times on chef and baskets, while ours changes by at most 7%. Down to 1 PPP, three orders of magnitude below full flux, our object depth MAE grows by at most 19% on any three-view scene (Tab. S3). Our Chamfer distance is lower on every three-view test scene and protocol flux level, but Transientangelo has better completeness and F-score at 2 mm in most of them (Sec. S4).

Lower photon counts. In Fig. 4 and Tab. S6, our mean object depth MAE stays below Transientangelo's at full flux, and below TV-L1 fusion and screened Poisson reconstruction [8] of our detected returns, down to fewer than 10^3 photons in three views. At the lowest flux, the reconstruction extends beyond the object, which the object depth MAE ignores (Sec. S5). Without masks, our object depth MAE at 0.01 PPP is 2.04 mm, against 1.94 mm with masks and 9.39 mm for the visual hull of the masks alone. Retrained with its official configuration, Transientangelo has a depth error of 5.48 mm on chef at 10 PPP, close to the 5.17 mm of its released checkpoint in Tab. 1. At 1 PPP, however, it recovers no surface with or without adapted settings (Sec. S3), whereas ours has 1.08 mm. With its photons pooled as in ours, the adapted Transientangelo recovers chef and baskets at 1 PPP with 7.6 and 1.9 times our depth error.

Ablation study. Table 2 compares our variants with TV-L1 fusion [49], the most accurate of three classical fusions down to 0.03 PPP (Tab. S6). Compared with NeRF's

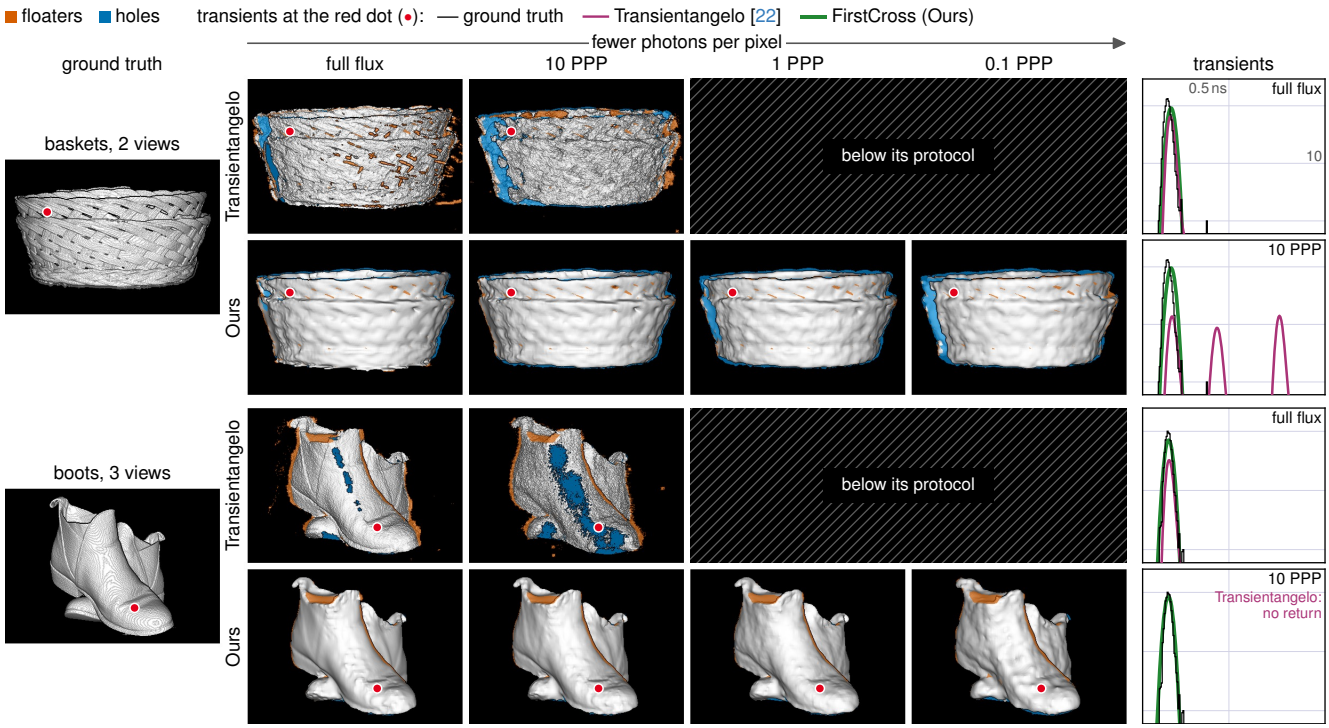


Figure 5. **Novel views from full flux to 0.1 PPP, and transients on two test scenes.** The red dot marks a pixel where Transientangelo has a floater or hole (depth error over 2 cm) at 10 PPP and both methods are within 5 mm at full flux. There, Transientangelo's transient at 10 PPP is mostly late on baskets and missing on boots, whereas ours stays at the ground-truth peak. Transients, on a log scale, are scaled to the ground truth (measured at full flux). Grey lines mark powers of ten (horizontal) and 0.5 ns after the ground-truth peak (vertical).

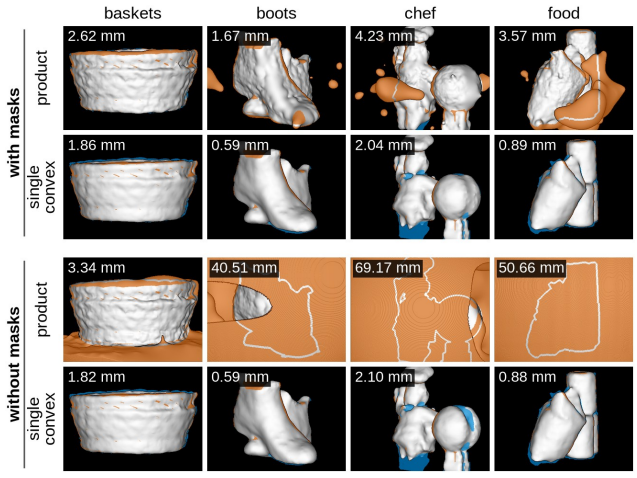


Figure 6. **Product transmittance against the single convex problem.** At 0.1 PPP on the three-view test scenes, the product variant leaves floaters (orange) with masks, largest on chef and food. Without masks, it covers most of the view of boots, chef and food with spurious surfaces. Blue marks holes, and numbers are object depth MAEs.

product transmittance [26], the closure helps most at low flux, without masks and when the solver starts from an occupied grid. Even the single convex problem cuts the

Table 2. **Ablation study.** Object depth MAE (mm, ↓) per flux level (PPP), averaged over the four three-view test scenes. Each variant changes one component and uses our solver. Product transmittance replaces the running maximum with NeRF's product. TV-L1 fusion uses our detected returns. Best in bold, second best underlined, and – marks settings not run.

	With masks			No masks
Method	Full	10	0.1	10
TV-L1 fusion [49]	1.18	1.61	2.03	–
Product transmittance	1.12	1.11	3.02	12.46
Single convex problem	0.91	0.92	1.35	0.91
FirstCross (Ours)	**0.90**	**0.89**	**1.17**	**0.85**

product variant's error at 0.1 PPP from 3.02 to 1.35 mm and removes most of its floaters (Fig. 6). From an empty grid, this problem has a lower object depth MAE than Transientangelo on every three-view test scene and protocol flux level (Tab. S15), and ours improves on it most at 0.1 PPP.

Without masks, the product variant produces spurious surfaces in front of the cameras (Tab. 2 and Fig. 6, Fig. S13), with 2.5 to 7.3 times our error at 0.01 PPP (Tab. S15). From the occupied grid $u = 0$, it leaves free space filled even with masks, whereas down to 0.1 PPP this start changes the single convex problem's error by at

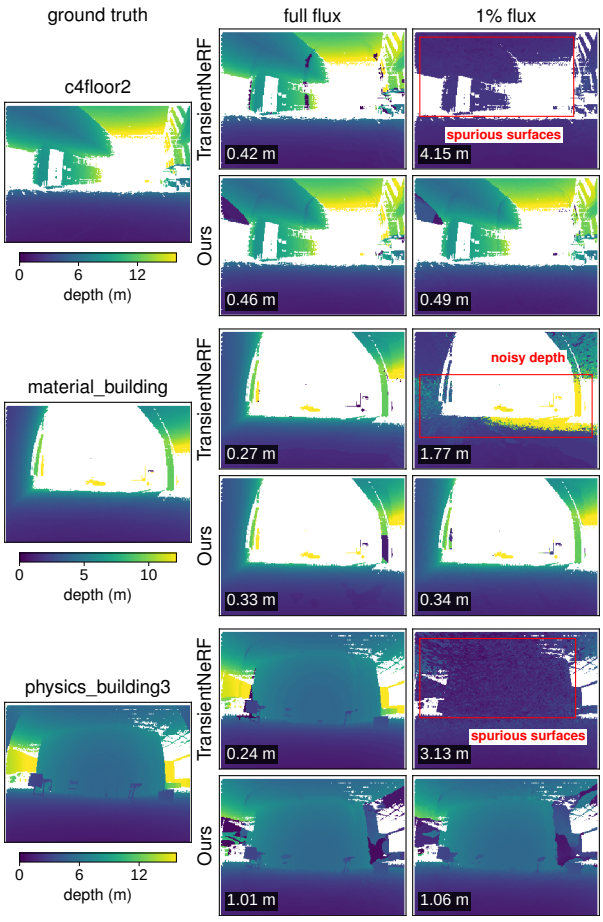


Figure 7. **Room-scale depth on the test rooms with three views.** At 1% flux, TransientNeRF's depth becomes noisy with spurious surfaces in every room (red frames), whereas ours stays near its full-flux depth. Per room, we show the novel view where our 1% flux error is the median, with one colour scale and white outside the benchmark mask. Numbers give the per-view mean depth MAE.

most 0.17 mm, due to inexact convergence. With L-BFGS from an empty grid, with or without masks, the product comes within 12% of the closure's mean error. From the occupied grid $u = \beta\varepsilon$, however, the product under L-BFGS fills free space without masks at 0.01 PPP on three of four scenes, unlike the closure (Tab. S16, Fig. S15). Without the too-early term, the error averages 4.6 to 40 times ours down to 0.1 PPP on validation and test scenes, and without the too-late term, holes open where photons are few. When q, τ or α is halved or doubled, the mean object depth MAE on the validation objects changes by at most 2%, and for ω or β, by at most 9% (Tab. S12).

Room scale. Across all rooms and view counts in Tab. 3, with unchanged dimensionless parameters, ours is 2.8 to 7.7 times more accurate than TransientNeRF at 1% flux (Fig. 7) and less accurate at full flux. We do not run Transientangelo on rooms, as its configuration assumes a bounding sphere

Table 3. **Mean depth MAE (m, ↓) over all novel views of the SP-TransientBench test rooms.** Better of each pair in bold. TN (TransientNeRF) trains for the official 300k steps. On material_building, our error drops at 1% flux as the grid box around our returns shrinks. With the full-flux box, ours wins there (Sec. S6).

		Full flux		1% flux	
Room	Views	TN [23]	Ours	TN	Ours
c4floor2	3	**0.435**	0.439	3.973	**0.525**
	5	**0.338**	0.459	3.891	**0.508**
material_building	3	**0.382**	0.705	2.135	**0.325**
	5	**0.319**	0.718	1.767	**0.315**
physics_building3	3	**0.768**	1.209	3.303	**1.193**
	5	**0.697**	1.066	3.455	**1.122**

of radius 0.4 m and object masks at low flux. Ours takes 3 min on library_floor2, whereas TransientNeRF trains for 10 to 17 h. On the test rooms, volumetric (TSDF) fusion [4] of our detected returns has 1.2 to 4.5 times our mean error and outperforms TransientNeRF at 1% flux (Sec. S6).

5. Conclusion

We presented FirstCross, which reconstructs solid objects from a few views of single-photon lidar when photons are scarce. For a solid object, the Poisson likelihood of a ray with given signal and background rates is a function of its first crossing alone. Its too-early part has a tight convex closure, volume rendering with a running maximum, as does the non-increasing part of any first-crossing cost. Its too-late part has a convex upper bound under the thickness assumption. FirstCross therefore works through a short sequence of convex problems without a neural network, in about a minute per object. It has lower depth error than Transientangelo [22] in every setting of the original real-capture protocol. With three-view inputs thinned 100-fold below the protocol's lowest flux, its mean object depth MAE stays under half of Transientangelo's at full flux. FirstCross suits few-view 3D scanning of objects and rooms when acquisition must be fast. Code will be released.

Limitations and Future Work. The thickness assumption requires objects thicker than the band ε, about 24 cm at room scale, and thin sheets break the first-iteration witness. Our field is also smooth at the scale of σ. At high flux, a neural surface is therefore sharper where views are dense, and on Transientangelo's three-view simulated scenes, our depth error is up to 1.36 times theirs above 10 PPP (Sec. S7). There, because our solid fills space that no training view sees, our Chamfer distance is up to 32 times theirs. On two validation scenes, adding ten times as many ambient photons as were recorded raises our error at 0.1 PPP up to tenfold (Sec. S5, which lists other limitations). Future work may refine the grid where views are dense, derive witnesses for thin sheets and estimate camera poses jointly. We also plan to extend the closure to other range sensors.

LLM Usage Statement

Claude (Anthropic) was used under the authors' direction, through the Claude Code agent, to write and review the experiment code, run and monitor the experiments, and compute the statistics. It also maintained an append-only experiment log and drafted the text and figures of this paper, including this statement. Claude's code generated our simulated low-photon inputs, the added ambient photons and the camera rotations (Secs. 4 and 5, supplement S5 and S7). The authors chose the research problem and the core method, approved each experiment plan before it was run, read every result, and edited the text. Gemini (Google) was used only to critique drafts. The authors take full responsibility for the content of this paper.

References

[1] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3), 2013. 3, 4

[2] Fatih Calakli and Gabriel Taubin. SSD: Smooth signed distance surface reconstruction. *Computer Graphics Forum*, 30 (7), 2011. 2

[3] Jonathan C. Carr, Richard K. Beatson, Jon B. Cherrie, Tim J. Mitchell, W. Richard Fright, Bruce C. McCallum, and Tim R. Evans. Reconstruction and representation of 3D objects with radial basis functions. In *Proc. SIGGRAPH*, 2001. 2

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH*, 1996. 2, 8

[5] Hongzhou Dong, Zili Zhang, Ziting Wen, Yiheng Qiang, Runrong Deng, Wenle Dong, Ziwen Jiang, Xinyang Li, Rui Lu, Shuoyao Sun, Wenyu Wang, Haitao Zheng, Guodong Shi, and Xiaoqiang Ren. SP-TransientBench: A real-captured single photon perception benchmark. In *Proc. European Conference on Computer Vision (ECCV)*, pages 101–118, 2026. 2, 6

[6] Felix Heide, Steven Diamond, David B. Lindell, and Gordon Wetzstein. Sub-picosecond photon-efficient 3D imaging using single-photon sensors. *Scientific Reports*, 8:17726, 2018. 2

[7] Fei Hou, Chiyu Wang, Wencheng Wang, Hong Qin, Chen Qian, and Ying He. Iterative Poisson surface reconstruction (iPSR) for unoriented points. *ACM Transactions on Graphics*, 41(4), 2022. 2

[8] Michael Kazhdan and Hugues Hoppe. Screened Poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3), 2013. 2, 6

[9] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proc. Eurographics Symposium on Geometry Processing*, 2006. 2

[10] Michael Kazhdan, Ming Chuang, Szymon Rusinkiewicz, and Hugues Hoppe. Poisson surface reconstruction with envelope constraints. *Computer Graphics Forum*, 39(5), 2020. 2

[11] Ahmed Kirmani, Dheera Venkatraman, Dongeek Shin, Andrea Colaço, Franco N. C. Wong, Jeffrey H. Shapiro, and Vivek K. Goyal. First-photon imaging. *Science*, 343(6166), 2014. 2

[12] Kalin Kolev, Maria Klodt, Thomas Brox, and Daniel Cremers. Continuous global optimization in multiview 3D reconstruction. *International Journal of Computer Vision*, 84 (1), 2009. 3

[13] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3), 2000. 3

[14] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Robust and efficient surface reconstruction from range data. *Computer Graphics Forum*, 28(8), 2009. 3

[15] Gabriel Lenain, Edmond Boyer, and Jean Ponce. Surface reconstruction from range scans under visibility constraints. *arXiv preprint arXiv:2610.18005*, 2026. 3

[16] Yue Li, Shida Sun, Yu Hong, Feihu Xu, and Zhiwei Xiong. 3D reconstruction from transient measurements with time-resolved transformer. *arXiv preprint arXiv:2510.09205*, 2025. 2

[17] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Siyou Lin, Dong Xiao, Zuoqiang Shi, and Bin Wang. Surface reconstruction from point clouds without normals by parametrizing the Gauss formula. *ACM Transactions on Graphics*, 42(2), 2023. 2

[19] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. Single-photon 3D imaging with deep sensor fusion. *ACM Transactions on Graphics*, 37(4), 2018. 2

[20] Shubao Liu and David B. Cooper. Ray Markov random fields for image-based 3D modeling: Model and efficient inference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1530–1537, 2010. 3

[21] László Lovász. Submodular functions and convexity. In *Mathematical Programming: The State of the Art*. Springer, 1983. 3, 4

[22] Weihan Luo, Anagh Malik, and David B. Lindell. Transient-tangelo: Few-viewpoint surface reconstruction using single-photon lidar. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 1, 2, 5, 6, 7, 8

[23] Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kiriakos N. Kutulakos, and David B. Lindell. Transient neural radiance fields for lidar view synthesis and 3D reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 6, 8

[24] Anagh Malik, Noah Juravsky, Ryan Po, Gordon Wetzstein, Kiriakos N. Kutulakos, and David B. Lindell. Flying with photons: Rendering novel views of propagating light. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2

[25] Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, and David B. Lindell. Neural inverse rendering

from propagating light. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 3, 7

[27] Bailey Miller, Hanyu Chen, Alice Lai, and Ioannis Gkioulekas. Objects as volumes: A stochastic geometry view of opaque solids. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[28] Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019. 5

[29] Fangzhou Mu, Carter Sifferman, Sacha Jungerman, Yiquan Li, Mark Han, Michael Gleicher, Mohit Gupta, and Yin Li. Towards 3D vision with low-cost single-photon cameras. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[30] Mukund Narasimhan and Jeff Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005. 3, 4

[31] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[32] Jiayong Peng, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. Photon-efficient 3D imaging with a non-local neural network. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[33] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable Poisson solver. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[34] Joshua Rapp and Vivek K. Goyal. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Transactions on Computational Imaging*, 3 (3), 2017. 2

[35] Nikolay Savinov, L'ubor Ladický, Christian Häne, and Marc Pollefeys. Discrete optimization of ray potentials for semantic 3D reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[36] Nikolay Savinov, Christian Häne, L'ubor Ladický, and Marc Pollefeys. Semantic 3D reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[37] Dominik Scheuble, Andrea Ramazzina, Hanno Holzhüter, Stefano Gasperini, Steven Peters, Federico Tombari, Mario Bijelic, and Felix Heide. Transient LASSO: Transient large-scale scene reconstruction. In *Proc. SIGGRAPH Asia Conference Papers*, 2025. 2

[38] Dongeek Shin, Ahmed Kirmani, Vivek K. Goyal, and Jeffrey H. Shapiro. Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors. *IEEE Transactions on Computational Imaging*, 1(2), 2015. 2

[39] Sébastien Speierer and Adrian Jarabo. Generalized non-exponential Gaussian splatting. *arXiv preprint arXiv:2603.02887*, 2026. 3

[40] Julián Tachella, Yoann Altmann, Nicolas Mellado, Aongus McCarthy, Rachael Tobin, Gerald S. Buller, Jean-Yves Tourneret, and Stephen McLaughlin. Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers. *Nature Communications*, 10:4984, 2019. 2

[41] Andrea Tagliasacchi, Matt Olson, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. VASE: Volume-aware surface evolution for surface reconstruction from incomplete point clouds. *Computer Graphics Forum*, 30(5), 2011. 3

[42] Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *Proc. International Conference on 3D Vision (3DV)*, 2015. 3

[43] Delio Vicini, Wenzel Jakob, and Anton Kaplanyan. A non-exponential transmittance model for volumetric scene representations. *ACM Transactions on Graphics*, 40(4), 2021. 3

[44] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[45] Ziting Wen, Runrong Deng, Zili Zhang, Haitao Zheng, Yuecong Xu, Xiaoqiang Ren, Guodong Shi, and Kemi Ding. Resolving mixed single-photon LiDAR returns for foreground-view and hidden scene reconstruction. *arXiv preprint arXiv:2610.01206*, 2026. 2

[46] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[47] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[48] Alan L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4), 2003. 4

[49] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2007. 2, 6, 7