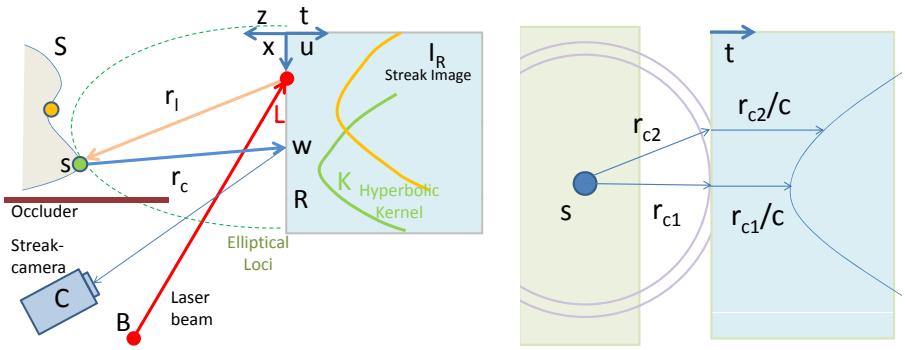
Recovering Three-dimensional Shape Around a Corner using Ultrafast Time-of-Flight Imaging

Supplementary Information

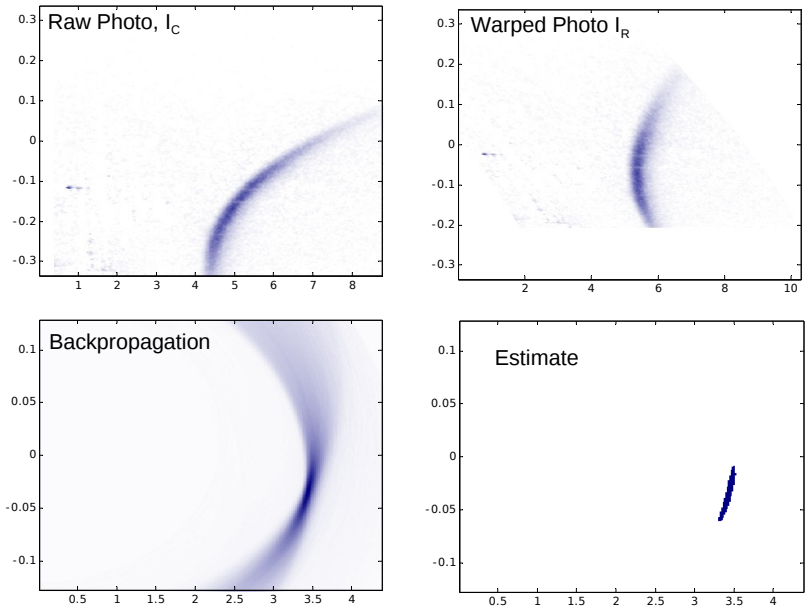
Andreas Velten, Thomas Willwacher, Otkrist Gupta,
Ashok Veeraraghavan, Moungi Bawendi, Ramesh Raskar

Supplementary Figures

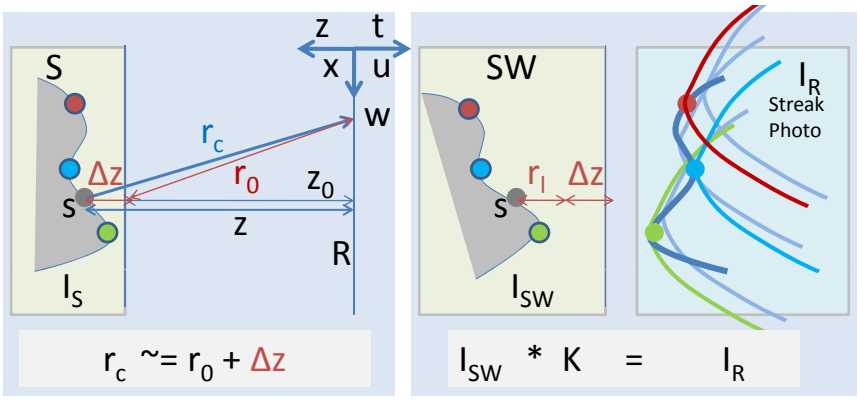


(a) (b)

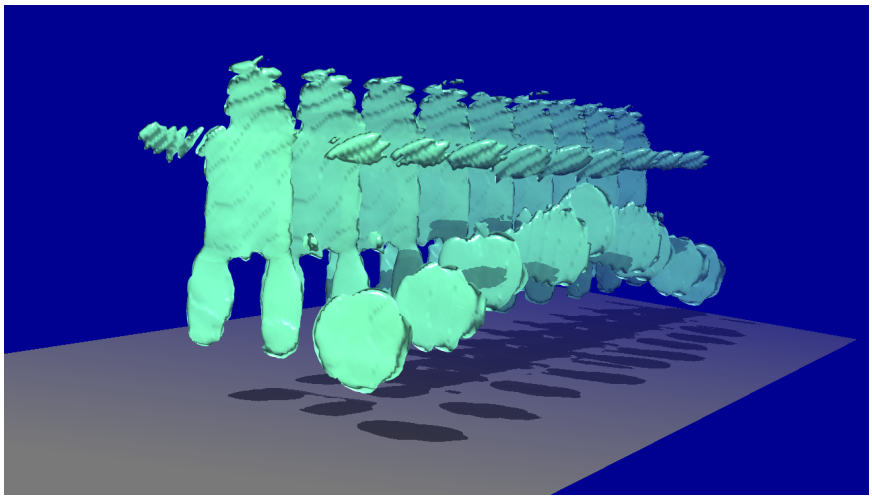
Supplementary Figure S1: Forward Model. (a) The laser illuminates the surface S and each point $s \in S$ generates a wavefront. The spherical wavefront contributes to a hyperbola in the space-time streak image, I_R. (b) Spherical wavefronts propagating from a point create a hyperbolic space-time curve in the streak image.



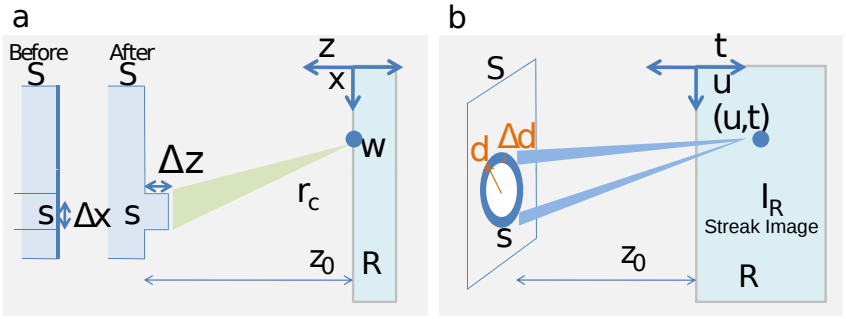
Supplementary Figure S2: Backprojection. A space time transform on a raw streak image allows us to convert a 4 segment problem into a sequence of 2 segment problems. The toy scene is a small 1cm$\times$1cm patch creating a prominent (blurred) hyperbola in the warped image. Backpropagation creates low frequency residual but simple thresholding recovers the patch geometry.



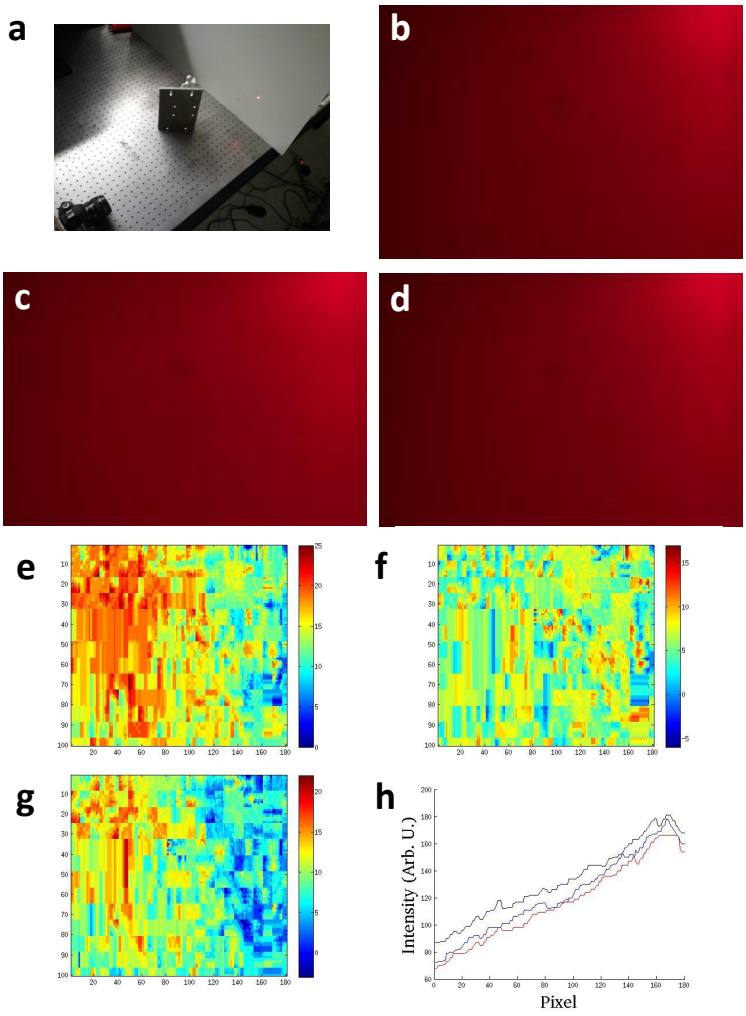
Supplementary Figure S3: Fresnel Approximation for convolution operation. With a near-constant depth assumption of $\Delta z \ll z_0$, the streak image I_R is approximated as a convolution of the *warped shape* countour image I_{SW} with the hyperbolically shaped kernel K. The warped shape image in turn is the true shape (S), deformed along the z direction according to laser distance. We assume an opaque object and hence the contributions are only from the points on the curve (surface) and not from the area behind the curve.



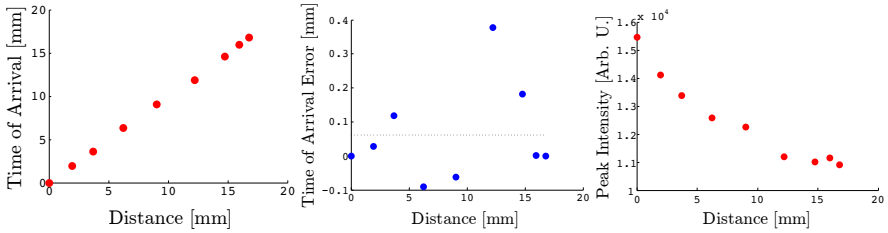
Supplementary Figure S4: Stop motion reconstruction. Results of a multi-pose stop motion animation dataset after filtered backprojection and soft-thresholding. A hidden model of a *man with a ball* is captured in various poses. The rendering shows the sequence of reconstructions created by our filtered backprojection algorithm and demonstrates the ability to remove low-frequency artifacts of backprojection. The mislabeled voxels remain consistent across different poses indicating stability of our capture and inversion process. Shadows are introduced to aid visualization.



Supplementary Figure S5: Improving depth and lateral resolution. (a) In a still camera, the ability to discern displacement of a patch with area $(\Delta x)^2$ by a distance Δz is limited by camera sensitivity. (b) Using time resolution, the ability to discern the same patch is improved and possible within practical camera sensitivity. The pixel (u, t) receives energy only from inside the ring. For simplicity, the diagrams in this document show the scene in flat-land and the surfaces are drawn as 2D curves.

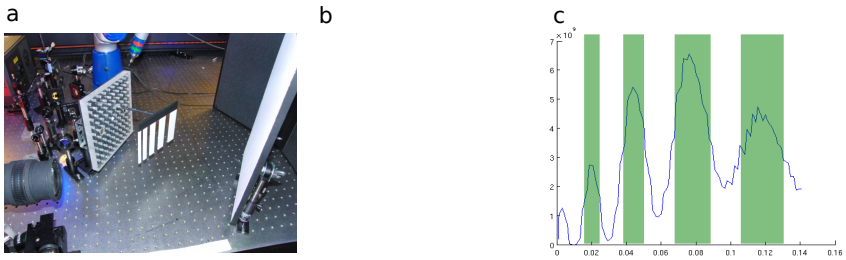


Supplementary Figure S6: Reconstruction attempt with a slow camera. We performed an experiment to demonstrate the challenges in imaging around the corner with a conventional, low temporal resolution laser and camera. (a) A setup with hidden mannequin but using a red continuous laser and a Canon 5D camera. (b) An image of the wall recorded with the Canon 5D camera with the room lights turned off and no hidden object present. (The recorded light is due to the reflections from walls behind the laser and camera.) (c) An image recorded with the hidden mannequin present. The increased light level on the wall is marginal, is low spatial frequency and shows no noticeable high frequency structure. (d) An image of the wall with the hidden mannequin moved away from the wall by 10 cm. The reduction in light level on the wall has no visible structure. (e) The difference between image in (b) and (c) using a false color map. (f) The difference between (b) and (d). (g) The difference between (c) and (d). (h) The plot of intensities along the centered horizontal scanline of each of the images (b=red, c=black, d=blue).



(a) (b) (c)

Supplementary Figure S7: Resolution in depth. (a) Distance estimation. Time here is measured in mm of traveled distance at the speed of light 1 mm≈0.3 ps. (b) Error is less than 1 mm. (c) Plot of intensity as a small patch is moved perpendicular to the first surface.



Supplementary Figure S8: Resolution in lateral dimension. (a) Setup with chirp pattern (occluder removed in this photo) (b) Raw streak photo from streak camera (c) The blue curve shows reconstruction of the geometry and indicates that we can recover features with 0.5 cm in lateral dimensions in the given scenario.

Supplementary Methods

Modelling Light Pulse Propagation

In this section, we analyze the relationship between the hidden scene geometry and the observed space time light transport in order to design methods to estimate the shape of the hidden objects. Consider a scene shown in Supplementary Figure S1. The scene contains a hidden object (whose surface we are interested in estimating) and a diffuser wall. A laser beam(B) emits a short light pulse and is pointed towards the diffuser wall to form a laser spot L. The light reflected by the diffuser wall reaches the hidden surface, is reflected and returns back to the diffuser wall. The streak camera is also pointed towards the wall.

For each location of the laser spot L, a $3D$ image (2 spatial and 1 temporal dimension) is recorded. The laser spot is moved to multiple locations on the wall ($2D$). The two dimensions for laser direction and the three dimensions for recording lead to a $5D$ light transport data. The pulse return time at each location on the wall depends upon several known parameters such as the location of the laser spot and unknown parameters such as the hidden surface profile. The idea is to exploit the observed $5D$ light transport data to infer the hidden surface profile.

For an intuitive understanding, consider the hidden scene to be a single point, as shown in Supplementary Figure S1. The reflected spherical wavefront propagating from that hidden scene point reaches the different points on the wall at different times creating a hyperbolic curve in the space-time streak image (Supplementary Figure S2). When the hidden scene contains a surface instead of individual and isolated scene points, the space-time hyperbolas corresponding to the different surface points are added together to produce the captured streak images and so we need to demultiplex or deconvolve these signals. In general, we could use a captured $5D$ light transport data but in our experiments, we are restricted to a streak camera that has a one spatial dimension. Thus, our system captures only a four dimensional light transport.

Bounce Reduction

In our setup, the optical path for light travel consists of 4 segments (Supplementary Figure S1): (1) from the laser B to the spot on the wall L, (2) from L to the scene point s, (3) from s again to a point on the wall w, and (4) finally from w to the camera C where it is recorded. However, the first and the fourth segment are directed segments and do not involve diffuse scattering. This allows us to precalibrate for these segments and effectively reduce the tertiary scattering problem to a primary (single) scattering problem. More concretely, suppose the camera records the streak image $I_C(p,t)$, where p is the pixel coordinate and t is time. In I_C, $t = 0$ corresponds to the instant the laser pulse is emitted from B. Then I_C is related to the intensity $I_R(w,t)$ of light incident on the receiver plane by the transformation

$$I_R(w,t) = I_C(H(w), t - ||L - B|| - ||C - w||). \tag{S1}$$

Here H is the projective transformation (homography) mapping coordinates on R to camera coordinates. The time shift by the distance from camera to screen, $||C - w||$, varies hyperbolically with the pixel coordinate w. Since the geometry of wall, R, is known, H, $||L-B||$ and $||C-w||$ can be computed in advance. Note there is no $\cos(\theta)$ factor or $1/r^2$ fall off in the above formula as the camera integrates over more pixels for oblique and distant patches. For this to hold, it is also important that R is Lambertian, as we assume. To summarize, the processing step (S1) reduces the problem to a single scattering problem, with an unfocused point source at L emitting a pulse at $t = 0$ and an unfocused virtual array of receivers on R recording the intensity of the reflected wavefront, $I_R(w, t)$.

Scattering of the light pulse

Generating Streak Images After the homography correction, we can consider a simplified scenarios of just two surfaces, the wall R and the hidden surface S. The surface S is illuminated by a light source at L. The surface R (receivers) can be assumed to host a virtual array of ultrafast photodetectors. The virtual photodetectors create an image $I_R(w, t)$ as intensity pattern of the incoming light as a function of time, t, and the position w. Hence the image, $I_R(w, t)$, is the intensity observed at $w \in R$ at time t. Experimentally, the virtual photodetectors are realized by using a Lambertian object R observed by a streak camera with ps time resolution (Supplementary Figure S1). Ignoring occlusions, the intensity pattern at R takes the following approximate form

$$I_R(w,t) = \int_S \int_\tau \frac{1}{\pi r_c^2} \delta(r_c - t + \tau) I_S(s,\tau) d\tau d^2 s \tag{S2}$$

where $w \in R$, $s \in S$, $t, \tau \in \mathbb{R}$ and $r_c = ||w - s||$. Furthermore, $I_S(s, \tau)$ encodes the hidden 3D shape S as the intensity of the light emitted by the transmitter at $s \in S$ at time τ. Note that we use units in which the speed of light $c = 1$. In other words, we measure time in units of distance. Note also that we make an approximation in neglecting the dependence on the normals to surfaces R and S. In the situation of interest to us, the object S is a diffuse (Lambertian) object illuminated by a single point source at position $L \in \mathbb{R}^3$. Concretely, this point source is the surface patch the laser is directed to. Hence, neglecting the normal dependence, $I_S(s, \tau) = I\delta(\tau - r_l)/(\pi r_l^2)$ with $r_l = ||L - s||$. Equation (S2) becomes

$$I_R(w,t) = \int_S I \frac{1}{\pi r_c^2} \frac{1}{\pi r_l^2} \delta(t - r_c - r_l) d^2 s \tag{S3}$$

The propagation of laser to wall and wall to camera is ignored in I_R. Laser to wall propagation is corrected using an offset value for time. The wall to camera sensor propagation is inverted by using a homography. In summary, the recorded streak image, I_C, which involves three or more bounces is converted to image, I_R, which involves only one bounce. For simplicity, we will ignore I_C and consider I_R as the streak image for rest of the discussion.

Hyperbolic Contribution For a fixed laser position, L, and sensor location, w, at a time t, the allowed values of s all lie on an ellipsoid with focal points L and w, given by the equation $t = r_c + r_l$. (More specifically, the locus of s lies on a *prolate spheroid*, i.e., an ellipsoid with two equal equatorial radii, smaller than the third equatorial radius.)

If we fix L and s this equation describes a two sheeted hyperboloid in (w, t)-space:

$$t - r_l = r_c = \sqrt{(x-u)^2 + (y-v)^2 + z(x,y)^2} \tag{S4}$$

where (u, v) are the two coordinates of w in the plane of the receiver wall. In particular, each point on the hidden surface S will contribute a hyperboloid to the image $I_R(u, v, t)$. The hyperboloids will have different shapes, depending on the depth $z(x, y)$, and will be shifted along the t-axis. Smaller depth $z(x, y)$ increases eccentricity and leads to higher curvature at the vertex of the hyperboloid .

Modified Fresnel Approximation Suppose that the hidden surface S has a small depth variation. We can write $z(x, y) = z_0 + \Delta z(x, y)$, with approximate mean depth z_0 and minor variations $\Delta z(x, y)$. Hence, $\Delta z(x, y) \ll z_0$. In this case, we apply an additional approximation, which is the analog of the Fresnel approximation in Fourier optics. Note that we are dealing with incoherent and pulsed light, so we call it the *modified Fresnel approximation*. Concretely, we expand the square root in (S4) and assume that $z_0 \gg (x - u)$ or $(y - v)$. The right hand side of (S4) becomes $r_c = \sqrt{(x-u)^2 + (y-v)^2 + z_0^2} + \Delta z(x, y)$, or $r_c = r_0 + \Delta z(x, y)$. Using this approximation in the argument of the delta function in (S3), and neglecting Δz in the denominator, we can express I_R as a convolution.

$$\begin{aligned} I_R(u, v, t) &\approx \int_{x,y} \frac{\delta(t - r_l - \Delta z - r_0)}{\pi^2 2 r_c^2 r_l^2} dx dy && \text{(S5)} \\ &= \int_{x,y,\tau} \frac{\delta(t - \tau - r_0)\delta(\tau - r_l - \Delta z)}{\pi^2 2 r_c^2 r_l^2} dx dy d\tau \\ &= (K * I_{SW})(u, v, t) && \text{(S6)} \end{aligned}$$

The hidden shape S is expressed using a delta function $I_S = \Delta z(x, y)$. Supplementary Figure S3 shows that, after a transform due to laser position, L, we have a new *warped shape approximation* $I_{SW}(x, y, \tau) = \delta(\tau - r_l - \Delta z(x, y))/(\pi r_l^2)$. We split the delta function inside the intergral above and re-write the equation as a convolution (in 3-dimensional (u, v, t)-space) of the warped shape approximation I_{SW}. This warped image I_{SW} "cramps up" information about the shape S in the time domain, warped by the additional "deformation" r_l, given by the distance to the laser. Finally the convolution kernel $K(x, y, t) = \delta(t - r_k)/(\pi r_k^2)$, with $r_k = \sqrt{x^2 + y^2 + z_0^2}$, is a hyperboloid, whose eccentricity (or curvature at the vertex) depends on z_0.

Note that equation (S6) is highly nonlinear in the unknown depths Δz, but linear in the warped shape I_{SW}, from which these depths can be determined. In conclusion, using

the modified Fresnel approximation, for every depth, we can express the forward propagation as a convolution with a hyperboloid. But for each depth, z_0, the curvature and position of the hyperboloid in space-time streak image, I_R, is progressively different.

Algorithms for surface reconstruction

Problem statement as a system of linear equations Let us express the results of the last section using linear algebra. Let us discretize the bounding box around the hidden shape and the corresponding 3D Cartesian space into voxels and arrange the voxels into a vector $f_S \in \mathbb{R}^N$. Here N is the number of voxels in the 3D volume of interest. The value of $f_S(i)$ is set to zero for all the voxels that do not contain a surface point $s \in S$. The value of $f_S(i)$ for voxels that lie on the surface of the objects is the albedo (or reflectance) of the corresponding surface point. Voxels that are interior to the object are occluded by voxels on the surface of the object and do not return any signal energy, so they are also set to zero. Consider now the streak image I_R recorded with the laser at position L_1. Vectorize the streak image pixels into a single vector $g_{R,1} \in \mathbb{R}^M$, where M is the total number of spatio-temporal pixels present. The pixel values will depend linearly on the albedos in f_S and hence satisfy a linear equation of the form

$$g_{R,1} = A_1 f_S, \tag{S7}$$

for some $M \times N$ matrix A_1. Concretely, the entries of A_1 can be read off from equations (S2) and (S3). If multiple streak images $1, \ldots, n$ are recorded corresponding to different locations of the laser, then those different streak images are stacked on top of each other in a vector y, which satisfies the linear equation

$$g_R = \begin{bmatrix} g_{R,1} \\ g_{R,2} \\ \vdots \\ g_{R,n} \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix} f_S = A f_S$$

Our goal is to analyze and solve the above linear system. The Fresnel approximation of the last section gives an intuition. The Fresnel approximation allows us to rewrite the linear system as $g_R = A_F f_{SW}$, where A_F is a block circulant matrix that represents the convolution with the hyperbolic kernel K.

Backprojection for surface reconstruction

Each voxel in the 3D world contributes signal energy to only a very small subset of the spatio-temporal bins (streak camera pixels) that are imaged. The specific spatio-temporal bins or pixels that contain signal energy are related to the hidden 3D surface. Further, if a particular pixel in the image contains no signal energy, this means that every voxel in the 3D world that would have contributed energy to it was empty. Both these pieces of information can be used to contruct an algorithm for reconstruction. The

basic intuition behind this algorithm (a backprojection algorithm, similar to algorithms used tomographic reconstruction) is very simple. Each observed pixel contributes energy back to all the source voxels that could have contributed energy to this pixel and the contribution made is proportional to the observed signal energy. Voxels in the world that were occupied receive contributions from all the pixels in the streak image that they contributed to and therefore have a large response. This response energy with appropriate filtering and thresholding can be used to recover the 3D surface. If the working volume is shallow (i.e., can be represented with a plane at depth z_0 plus minor depth variations), we can use a single kernel of Frensel approximation leading to a block circulant matrix A. Below, we first explore reconstruction with Frensel approximation. Then we strive for a more accurate result. We use a different kernel for each depth, leading to a non-block circulant matrix.

Depth Independent Fresnel Approximation In order to implement the backpropagation algorithm, in practice, it is necessary to model the forward propagation of the spherical wavefronts from each of the hidden surface voxels. Although approximate, we first use the Fresnel approximation based forward propagation model described in the Section "Scattering of the light pulse" for a better understanding and ability to easily analyze the invertibility. Under the Fresnel approximation the captured streak images can be written as a convolution of the unknown surface with the hyperbolic blur kernel as given by

$$I_R = K * I_{SW}$$

with $K(x, y, t) = \delta(t - \sqrt{x^2 + y^2 + z_0^2})/(\pi(x^2 + y^2 + z_0^2))$. The backprojection kernel, on the other hand, is $\tilde{K}(x, t) = \delta(t + \sqrt{x^2 + y^2 + z_0^2})$. Hence backprojection is, up to the distance attenuation prefactor, the adjoint of propagation. If the Fresnel approximation is valid, the effect of backprojection on the captured streak images can be described as a convolution with the point spread function $psf = \tilde{K} * K$. The function psf can be computed analytically. This function has a large peak at the center, surrounded by a butterfly shaped low frequency component. This peak implies that when one performs backprojection peaks will be observed at all the locations where there is a 3D scene point. The limitations of backprojection are also evident from the function psf. Since the peak is surrounded by low frequency components, this approach without any post-processing (filtering) will lead to overly smoothened results. Rephrasing these observations in the matrix notation introduced at the beginning of this section, one can say that the backprojection operation is described by the matrix $\tilde{A}_F$, which is the same as A_F^T,up to the distance attenuation factors. The composition of propagation and backprojection $\tilde{A}_F A_F$ is close to the identity matrix.

Depth Dependent backprojection While it is very useful to use the Fresnel approximation to analyse the effect of backprojection, the approximations made lead to inaccuracies when (a) the scene has significant depth variation or (b) there are occlusions. In those cases we need to use the more precise formulas (S2), (S3). Propagation can then no longer be written as a convolution, since the integral kernel, i.e., the hyperbola, changes shape with varying depth.

Limitations of Backprojection Backprojection suffers from several limitations. The results of backprojecton are smoother than the original surface and there are still a few false positive surface points. We also observe that surface slopes beyond 45 degrees are extremely hard to reconstruct. This can be explained theoretically, at least in the Fresnel approximation as follows. The Fourier transform of the hyperbolic convolution kernel falls off gently (by a power law) in the direction orthogonal to the receiver plane, but falls off exponentially in the parallel direction. Hence features having high parallel spatial frequencies, such as high slope regions, are very hard to recover. In order to tackle these limitations it would be necessary to use additional prior information about the 3D scene being reconstructed.

We use the traditional approach that removes the low frequency artifacts around sharp features. The method is known as filtered backprojection that involves using a carefully chosen high pass filter. In our case, a good choice is second derivative in the z direction of the voxel grid. Supplementary Figure S4 shows the results after such filtering and applying a soft threshold described in the main paper.

Note that backprojection is a voxel-based technique and does not take into account the surface-based properties like orientation or reflectance profile. Hence our technique is expected to work best for nearly Lambertian surfaces, for which the recorded images do not depend strongly on the surface normals.

Necessity of ultrafast imaging

We consider the achievable resolution and space-time dimension tradeoffs in hidden shape recovery.

Limits of Traditional Photography

Even with a still camera, one can in principle detect the displacement of a small hidden area of a scene as shown in Supplementary Figure S5(a), but the problem is ill-conditioned. To see this, let us consider for simplicity a near planar scene at depth z_0, illuminated homogeneously by a far away source. The intensity of light incident at a point $r \in R$ that was emitted by a surface patch above r of size Δx by Δx is proportional to $I(\Delta x)^2/(z_0^2)$, where I is the total intensity received. Moving the patch by $\Delta z \ll z_0$ in depth, the contributed intensity will change by $\Delta I \propto I(\Delta x)^2 \Delta z / z_0^3$. Hence we conclude that $\Delta I/I \propto (\Delta x)^2 \Delta z / z_0^3$. As in typical scenario, the spatial resolutions ($\Delta x, \Delta z \approx 5$mm, $z_0 \approx 20$cm, we require intensity resolution, $\Delta I/I \sim 3 \times 10^{-5}$. This means one has to distinguish intensities of 1 from 1.00003. This is not possible in practice. Note that the intensities received after tertiary scattering are already very small, so it is hard to obtain a good signal to noise ratio. We show the limits of traditional low temporal resolution photography via an example in Supplementary Figure S6.

Benefits of Time Resolution

For an ordinary camera, two conditions make the problem ill-conditioned: The relative intensity contributed by an emitter changes only slightly ($\propto \Delta z/z_0$) and this small change is overwhelmed by the contribution of the background with area A, yielding the factor $(\Delta x/z_0)^2$. Using a ps accurate high-speed light source and sensors these problems can be circumvented.

A change in patch position s means it contributes to a different pixel in the streak photo, provided $\Delta z/c > \Delta t$, where $c =$ speed of light and Δt is time resolution.

Unlike an ordinary sensor, not all patches on S contribute to a pixel (time bin) in a streak photo making the mixing easier to invert. The locus of points contributing to a fixed sensor and time-bin position, (u, t), lie on a ring with radius $d = \sqrt{(ct)^2 - z_0^2}$ (Supplementary Figure S5(b)). If the time bin has a width of $\Delta t \ll t$, the width of the ring is approximately $\Delta d = c^2 t \Delta t/d$. Hence the total area of the ring is $2\pi d\Delta d = 2\pi c^2 t \Delta t$. We want to detect changes in the intensity emitted by a patch of size $\Delta A = (\Delta x)^2$. Hence the change in total intensity is approximately $\Delta I/I = \Delta A/(2\pi d\Delta d) = (\Delta x)^2/(2\pi c^2 t \Delta t)$. In our scenario typically $\Delta x \approx 3c\Delta t$. Furthermore $ct \approx z_0$. Hence $\Delta I/I \approx 3\Delta x/(2\pi z_0)$. Thus the required intensity increment is linearly proportional to $\Delta x/z_0$, and not quadratically as before. In our case, this ratio is a reasonable $\sim 10^{-2}$. This gives the guidance on time-resolution. In addition, the time resolution of the light source should not be worse than that of the sensor.

Performance Validation

We performed a series of tests to estimate the spatial resolution perpendicular and parallel to the visible surface, i.e., the wall. We use the FARO Gauge measurement arm to collect independently verifyable geometric position data (ground truth) and compared with positions recovered after multiple scattering using our algorithm. In our system, translation along the direction perpendicular to the diffuser wall can be resolved with a resolution of 400 μm better than the full width half maximum (FWHM) time resolution of the imaging system (Supplementary Figure S7, a and b). Lateral resolution in a plane parallel to the wall is lower and is limited to 0.5–1 cm depending on proximity to the wall (Supplementary Figure S8).

Choice of resolutions

There are four important parameters for our shape estimation setup: The spatial resolution, i.e., the spacing between sensor pixels, the temporal resolution of sensors and the laser, the intensity resolution and signal to noise ratio of our sensors (and power of the light source) and the angular diversity determined by the geometry of the setup. We saw in Section "Benefits of Time Resolution" that time resolution is critical and gives us an approximate lower bound on the resolution of our 3D reconstruction This is the same as in the case of a direct view traditional time of flight camera. However,

our situation differs from a direct view time of flight camera. Neither our sensors nor our light source have any directional resolution into the hidden space after a diffuse reflection.

Spatial Camera Resolution If we could determine the correspondences, i.e., which part of the received signal at a sensor was emitted by which transmitter (surface patch), spatial resolution of a streak camera would actually be unnecessary. Time resolution will directly determine reconstructed 3D resolution of hidden objects. Despite these two challenges, finite time resolution of the streak camera and the loss of correspondence, the sufficiently high spatial resolution allows us to exploit the local structure in streak photo to recover shape without explicitly solving the correspondence problem.