Supplementary Material: Occluded Imaging with Time of Flight Cameras

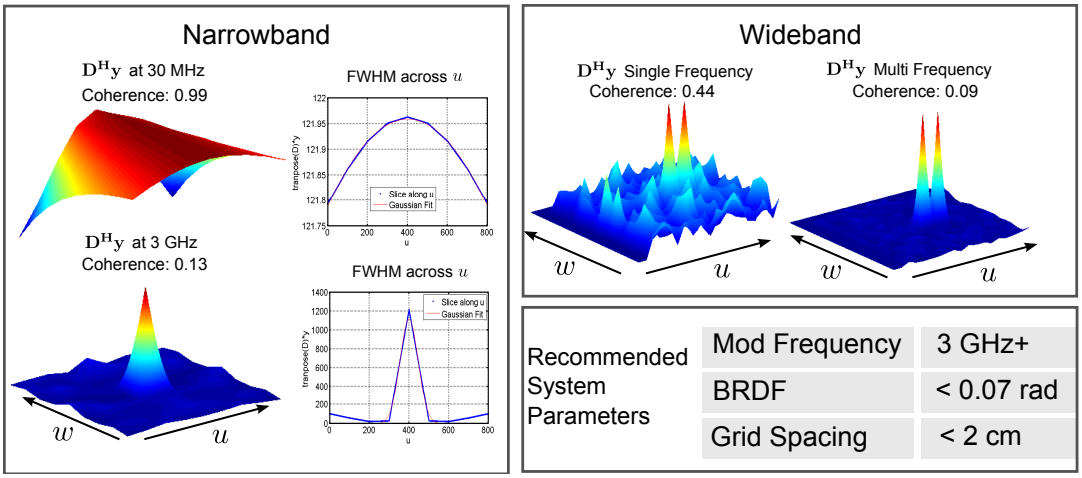


Figure 1: *Impact of modulation frequency on results. In the narrowband case increasing the frequency serves to decrease the width of the beampattern. Given the same modulation frequency; wideband data helps to mitigate spatial aliasing artifacts*

1 Extended Recovery Guarantees

In the main paper, we provided the derivation for specularity and its effect on mutual coherence. Here, we perform a similar analysis of other parameters using constraints on rank, span, and mutual coherence.

1.1 Span Constraint

Suppose there are M virtual sensors that lie within P dimensional space. The m-th virtual sensor is associated with a location in P dimensional space represented as a vector $v_m \in \mathbb{R}^P$. The location vectors for all virtual sensors, is denoted as the set $\{\vec{v}_1, \ldots, \vec{v}_M\} \in \mathbb{R}^P$. From the location of the virtual sensors, we can define the subset $V = \left\{\vec{v}_1 + \text{span}\left\{\vec{v}_i - \vec{v}_j\right\}_{i \neq j}\right\}$, shown geometrically in Figure 2. Now, suppose there are N target locations within P dimensional space, with location vectors $\left\{\vec{t}_1, \ldots, \vec{t}_N\right\} \in \mathbb{R}^P$. Denote $W = \left\{\vec{t}_1, \ldots, \vec{t}_N\right\}$ to represent the set of target location vectors.

Recall that the proposed recovery technique relies on inverting the complex matrix $\mathbf{D}$ as described in the paper. Without loss of generality, consider unique localization of a single source using only time-of-flight information. Since distance equals the product of time-of-flight and the speed of light, $\mathbf{D}$ can be written as a scaling of the well-known Euclidean Distance Matrix:

$$\mathbf{D} = c \begin{bmatrix} d_{11} & \cdots & d_{1N} \\ \vdots & & \vdots \\ d_{M1} & \cdots & d_{MN} \end{bmatrix}, \quad c = 3 \times 10^8 m/s,$$

where d_{mn} represents the Euclidean distance between the m-th virtual sensor and the n-th target location, written as $\|v_m - t_n\|_2$. A unique representation fails to exist when two targets have the same distance wrt. the virtual sensors (i.e. for targets $\vec{t}_p$ and $\vec{t}_q$, $\|\vec{t}_p - \vec{v}_m\|_2 = \|\vec{t}_q - \vec{v}_m\|_2$, for $m = 1, \ldots, M$). In other words, recovery of a single source will fail if two columns of $\mathbf{D}$ are identical.

Claim: If $W \not\subseteq V$, i.e, the target locations are not within the virtual sensor span, then unique recovery is not guaranteed.

Proof: Since $W \not\subseteq V$, for each $m = 1, \ldots, M$, there exists a target vector $\vec{t}_p \notin V$, with residual projection $\vec{r}_m = \vec{t}_p - \vec{v}_m \notin V$, where $\vec{r}_m \neq 0$. Suppose there exists another target vector $\vec{t}_q \notin V$, such that $\vec{t}_q = \vec{v}_m + (-\vec{r}_m)$. Since $\vec{v}_p \neq \vec{v}_q$, and $\|\vec{t}_p - \vec{v}_m\|_2 = \|\vec{t}_q - \vec{v}_m\|_2 = \|\vec{r}_m\|_2$, for $m = 1, \ldots, M$, recovery is not necessarily unique.

In summary, if $W \not\subseteq V$, degenerate configurations (fundamental problems with unique representations) can occur. This is emphasized geometrically in Figure 2 of this document. Please see the caption for details.

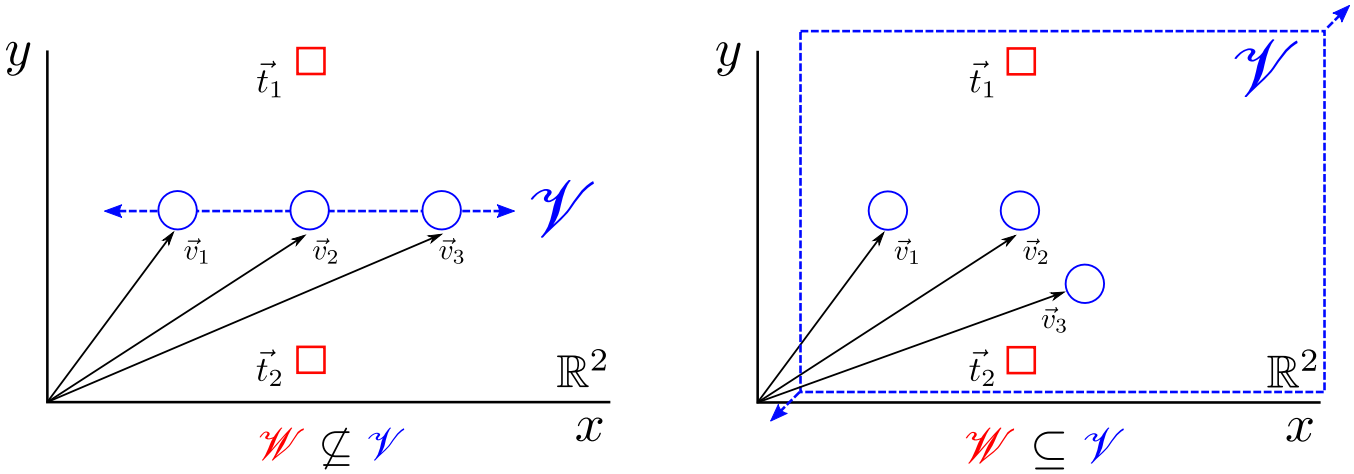


Figure 2: *A geometric interpretation of the array span constraint. (a) For a 1D, linear array, V is a 1-dimensional subspace. If the set of target location vectors span a 2 dimensional space, degenerate configurations occur, such as this case where $\vec{t}_1$ cannot be distinguished from $\vec{t}_2$. (b) By changing the configuration of the sensor array, V now spans $\mathbb{R}^2$, and for targets within $\mathbb{R}^2$, degeneracies are avoided.*

1.2 Rank Constraint

For this section, assume that $\mathbf{D}$ takes the form of a canonical Euclidean Distance Matrix. The *rank-constraint* is given as $\text{rank}(\mathbf{D}) - 2$ and encodes an upper bound on the dimensionality of the convex hull of targets (cf. pages 81-97 of [Gower 1985]). In practical settings, this is used in the multilateration problem (central to GPS), specifying the minimum number of satellites required to localize a target in N-dimensional space. Matrix rank is a black and white phenomena. The matrix $\mathbf{D}$ might satisfy the rank constraint, but the columns may be very close to linearly dependent. The next section discusses the metric of mutual coherence, which provides a convenient framework to provide physical guarantees.

1.3 Mutual Coherence

FWHM and coherence Because we are dealing with Gaussians there is a relationship between the FWHM of the Gaussian and mutual coherence. First, consider the beampattern function (i.e. the first row of $\mathbf{G}$). This function is Gaussian.[1] From the definition of $\mathbb{G}$, we know two points on the Gaussian (the "'center"' is at 1, and the closest point to the center is the mutual coherence). Since we have two points on the Gaussian, and we know the mean, the variance and FWHM can be found (note that for a normal distribution, $\text{FWHM} = 2\sqrt{2\ln 2\sigma}$).

Frequency and gridding We now derive a closed form equation to relate grid spacing and frequency to coherence. Suppose i' and j' represent the indices of the columns of $\mathbf{D}$ that determine the mutual coherence, where $\mathbf{D}_{i'} = \begin{bmatrix} \exp(-j\varphi_{i',1}) & \cdots & \exp(-j\varphi_{i',1}) \end{bmatrix}^\top$. Then the mutual coherence is $\mu(\mathbf{D}) = \sum_{m=1}^{M} \exp\left(-j\left(\varphi_{i',m} - \varphi_{j',m}\right)\right)$. Substitution using Equation 1 from the main paper yields:

$$\mu(\mathbf{D}) = \left| \sum_{m=1}^{M} \exp\left(-j\left(\frac{2\pi}{c} f_{\mathcal{M}}\left(z_{i',m} - z_{j',m}\right)\right)\right)\right| \tag{1}$$

where $z_{i',m}$ and $z_{j',m}$ represent the propagation distances for the i-th and j-th voxels to the m-th sensor. The main implication is as follows: increasing $f_{\mathcal{M}}$ or decreasing the resolution of the grid would decrease coherence. Also note that both grid spacing and modulation frequency are linear with respect to the exponent.

Aperture size and camera resolution In Equations 20 and 21 of the main paper, the aperture size D is inversely proportional to resolution. This suggests that, when possible it is beneficial to use larger objects as virtual sensors. Another interesting question is how many sensors are used? It turns out that using more or less sensors does not directly change the resolution of targets. However, using too few sensors leads to spatial aliasing artifacts, especially at high modulation frequencies. In particular, sensors must be spaced apart a distance of λ to avoid spatial aliasing.

For this problem, the camera's spatial resolution determines how many pixels are mapped to the virtual sensor array. At typical modulation frequencies, aliasing will then not be a factor.

Sparsity of the scene Theorem 1 suggests that recovery becomes harder when the scene is dense (K is high). For dense scenes, Theorem 1 is obviated by the bound in Equation 21 of the main paper. The interesting question is whether sparsity should be factored into the

[1] Following from, simplifications from computer graphics this can be approximated as a Gaussian.

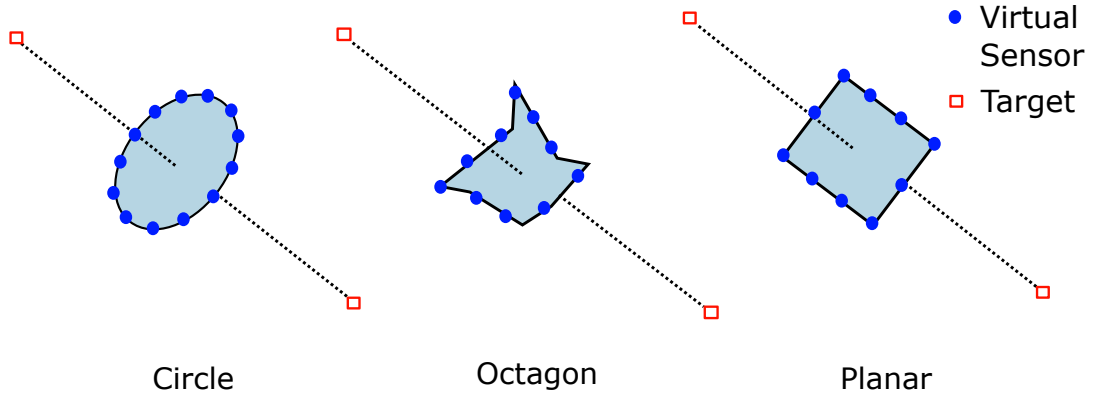


Figure 3: *Geometry of the virtual sensor array is governed by the span constraint. Symmetric degeneracies occur if the spanning constraint is not met, regardless of sensor geometry. Here, three different geometries each exhibit degeneracies.*

reconstruction. If this was a conventional camera taking regular pictures, it is clear that the scene is not canonically sparse. However, in the corners problem, obtaining even a sparse reconstruction might be useful. For example, if the goal was to localize occluded human targets for military applications. In such cases imposing a sparse prior can lead to more stable recovery.

Recovery algorithms We now turn to the choice of recovery algorithm. If the scene is not sparse then reconstruction via backprojection or pseudoinverse is optimal in the sense of minimizing the ℓ_2 error. If the scene is sparse then either a greedy algorithm, e.g. compressive sampling matching pursuit (CoSaMP), or a convex relaxation, e.g. Basis Pursuit Denoise, will result in better performance.[2] Recovery with sparse priors will allow, in practice, for recovery of closely spaced sources beyond the Rayleigh limit, i.e., superresolution (see Section 6 of the main paper). Table 1 of the main paper details the recovery algorithms that we use in experiments [3]. If the scene is not sufficiently sparse, then recovery is no longer guaranteed. In this paper, we tend to focus on backprojection or pseudoinverse solvers because: (i) they are very easily implemented and (ii) are much more general.

Geometry of virtual sensor array: The placement of virtual sensors impacts recovery. We begin with a failure case. Suppose an array of virtual sensors are placed on the circumference of a circle. Dictionary atoms corresponding to two sources placed on the axis will be linearly dependent, and therefore recovery will fail. To avoid such situations, the candidate target locations must lie within the span of the virtual sensor locations. The failure case violates this condition, as the sensors span a plane that does not include axial target locations. This is shown visually in Figure 3.

1.4 Simulated Validations

Similar analysis follows for modulation frequency and grid spacing. From Equation 1 it is expected that both a higher modulation frequency ($f_{\mathcal{M}}$) or spacing between grid points (which increases $z_{i',m} - z_{j',m}$) would lead to decreased coherence. Because both $f_{\mathcal{M}}$ and $z_{i',m} - z_{j',m}$ are linear w.r.t. exponent and they should have a similar impact on coherence. Finally, we expect this relationship to be approximately linear (to see why, use a small angle approximation for Equation 20 from the main paper). Now we turn to the simulation, where the results verify the analysis in that both modulation frequency (cf. Figure 5c) and grid spacing (cf. Figure 5d) have a linear relationship to coherence.

Two important consequences follow from this result. Doubling the modulation frequency, $f_{\mathcal{M}}$, would double coherence and thus resolution. In contrast, but reducing γ until the curve falls into the linear region is much more critical for minimizing coherence.

2 Reconstruction Noise Limit

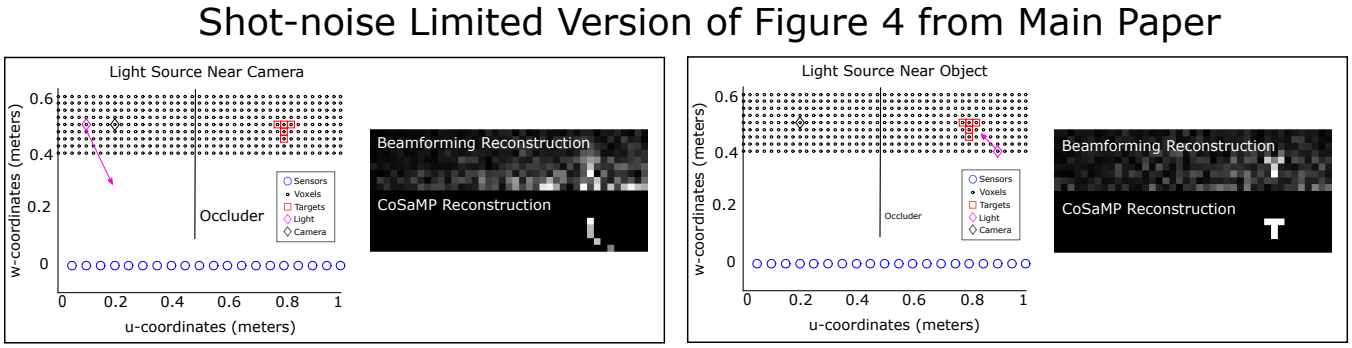
Intuitively, it may seem that the extra bounces in Figure 4a of the main document may lead to a worse reconstruction as compared to Figure 4b. However, this is not necessarily true, which we will show through a noise-limiting analysis. Specifically, there are two primary factors that contribute to the noisy pattern observed in the reconstruction: (i) camera noise statistics, encoded in the measurement vector $\vec{y}$, and (ii) reconstruction "noise".[4]

Our simulation in Figure 4 of the main paper models camera noise by considering three additive components of noise (shot, read, and dark noise). As is typical, we ignore dark noise (since it is small, and can be further reduced by cooling the sensor). Then, the measurements can then be characterized as either read-noise limited and shot-noise limited. The former occurs when few photons are measured, such that the variance from read-noise is much greater than that of shot-noise (i.e., $\sigma^2 = \sigma^2_{shot} + \sigma^2_{read} \approx \sigma^2_{read}$). Similarly, in the latter case many photons are measured (e.g. a long exposure), such that the shot-noise variance dominates the signal ($\sigma^2 = \sigma^2_{shot} + \sigma^2_{read} \approx \sigma^2_{shot}$).

[2] Convex relaxations are more robust than greedy algorithms, while the simplicity of greedy algorithms facilitates model based recovery.

[3] For details on CoSaMP and Basis Pursuit Denoising see [Needell and Tropp 2009] and [Chen et al. 1998].

[4] What is "noise"? Quotes are used because some communities use a broad characterization of noise to encompass what is not signal, while other communities impose a necessary condition of characterizing or modelling noise as random fluctuations. We are using the former convention.



(a) Light on Camera Side

(b) Light on Object Side

Figure 4: *Same as Figure 4 from main paper, but using less light, so shot-noise limits the spatial reconstruction, instead of inherent reconstruction noise. Please see Section 2 of this document for details. Not shown: if less light is used, the reconstruction in (a) becomes worse, while (b) will stay in a reconstruction-noise limited case. But as the light is further decreased, both (a) and (b) will be shot-noise limited.*

However, because we are not simply taking a photograph, there is another "noise" element that needs to be considered. This is reconstruction-noise, which characterizes the unwanted artifacts observed in the spatial pattern of the reconstruction due to mutual coherence effects. Because reconstruction-noise is not necessarily random, this limiting case is defined quite differently from read and shot-noise limits. We use the following definition:

Definition: *A system is reconstruction-noise limited if the signal-to-noise ratio of the measurement vector* $SNR(\vec{y}) > c\frac{1-\mu(\mathbf{D})}{\mu(\mathbf{D})}$*, for some constant* $c > 0$*.*

To explain the comparison of Figure 4a and 4b of the main document, we first note that the source intensity is the same for both models. However, because the source intensity is high, the SNR after incorporating camera-noise statistics alone is high.[5] In this case, reconstruction-noise contributes to the observed spatial noise of the reconstruction. Since the reconstruction noise depends primarily on the mutual coherence of the scene setup—refer to the second to last paragraph of section 4.3 of the main paper—it is expected that Figure 4a, and 4b show similar quality reconstructions.

To ensure that such a model is valid for varying scene conditions, a second experiment was performed using the same simulation model, where the intensity of the light source was lowered (see Figure 4 of this document). Now, the intensity is low-enough, such that the system is limited by signal-dependent shot noise. Then, as expected, this leads to the case that Referee 2 suggests. The scenario in Figure 4a is reconstructed with less fidelity than 4b of the supplement (since the received signal amplitude of Figure 4a is much lower due to a second attenuation by the wall and additional falloff from the inverse square law).[6]

3 Virtual Sensor Directionality: Experimental Methods

We measure the virtual sensor directionality of three typical walls using the ToF measurement. The mirror is skipped since its response is simply a delta function. The directionality is only determined by the BRDF of wall if the light is an ideal omnidirectional point source (as shown in the synthetic test on MERL BRDF in Figure 11). In our real device, the LED emits light in a non-uniform manner, which contains some unknown intensity distribution, which also shows some directionality. The virtual sensor directionality we are going to measure and fit is the joint effect of directionality of wall and light, which jointly determines the localization resolution (or FWHM) in the real system.

We first calibrate the extrinsic and intrinsic parameters of our ToF camera by using the camera calibration toolbox in Matlab. Then we put the light source and camera as close as possible and keep them a fixed distance to the wall (1m in our test). We capture an amplitude image of each wall with this setup, and measure all relative distances from light source and camera location to all pixels the image. We consider the inverse square law for the amplitude decay, since the light is relatively close to the target surface. The light transport and image formation has been described in Equation 5-7 of the main paper. Here we focus on the amplitude image with all distance calibrated to remove the phase part.

We empirically found that one dominant lobe could be observed in our amplitude images, and an isotropic Ward lobe [Ward 1992] is accurate enough to fit our observation. The specular lobe of the Ward model is given in Equation 2.

$$S(\rho_s, \alpha) = \frac{\rho_s}{4\pi\alpha^2\sqrt{(\mathbf{n}^\top\mathbf{l})(\mathbf{n}^\top\mathbf{v})}} \exp\left(\frac{\left(1-\frac{1}{\mathbf{n}^\top\mathbf{h}}\right)}{\alpha^2}\right), \tag{2}$$

where $\mathbf{n}$, $\mathbf{l}$, $\mathbf{v}$ are unit vectors of surface normal, incident lighting, and viewing directions; $\mathbf{h}$ is the half vector as $\mathbf{h} = (\mathbf{l}+\mathbf{v})/\|\mathbf{l}+\mathbf{v}\|$, which is the bisector of $\mathbf{l}$ and $\mathbf{v}$. We denote the incident angle θ_l as the angle between $\mathbf{n}$ and $\mathbf{l}$, and the exitant angle θ_v as the angle between $\mathbf{n}$ and $\mathbf{v}$. All these directional vectors can be calculated given the calibrated relative distances. With these vector directions, we fit Equation 2 to

[5] The source was modeled as a 2W laser source, which is a class IV laser.

[6] Reflective surfaces (wall and object) modeled with an albedo of 0.1.

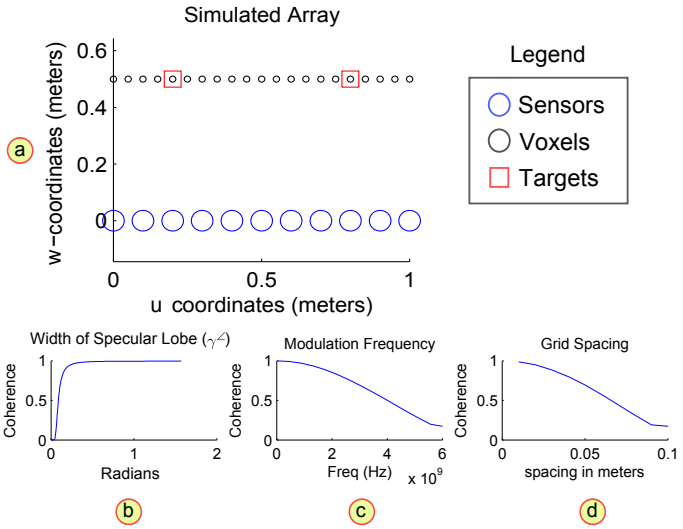


Figure 5: *(a) The simulated array used. (b-d) Impact of physical parameters on coherence. This figure is the exact same as Figure 10 from the main paper. It is redrawn here to make the supplement self-contained.*

all measured amplitude values. The fitting is calculated by solving a nonlinear least squares problem, using the Matlab function "lsqnonlin". The fitted parameter (ρ_s and α) and rendered directional response is shown in Figure 11 of the main paper, with response parameters in Table II. For a given incident angle ($\theta_l = 30^\circ$ in this example), the FWHM$^\circ$ (degree) is evaluated by plotting the fitted values for all exitant angles $\theta_v \in (-\frac{\pi}{2}, \frac{\pi}{2}]$.

4 Multipath Propagation

Multipath propagation is a very technical detail for AMCW cameras. The basic idea is that a single pixel receives photons from multiple depths. Because AMCW cameras use coherent illumination, this results in a problem of unicity. In this paper, we use spatial diversity to implicitly address the multipath problem. In other words, each virtual sensor array element measures different linear weightings of the transmitting sources which allows us to implicitly invert in the linear inverse problem $y = Dx$. For details on multipath propagation, please refer to

5 Narrowband vs Wideband

One of the differences between our method and that of Velten et al. and Heide et al. is that our technique does not require wideband (multifrequency) data. We can even make a stronger statement—using our method wideband data does not provide any benefit in resolvability. This is well known in the signal processing community where adding wideband data does not provide increased resolution over using the highest narrowband (single frequency) case. We now verify this claim with a simulation. As illustrated in Figure 1, a higher narrowband frequency yields greater resolvability. However, for the same narrowband system, adding extra frequencies of lower value do not improve resolution, though they do serve to eliminate spatial aliasing artifacts.

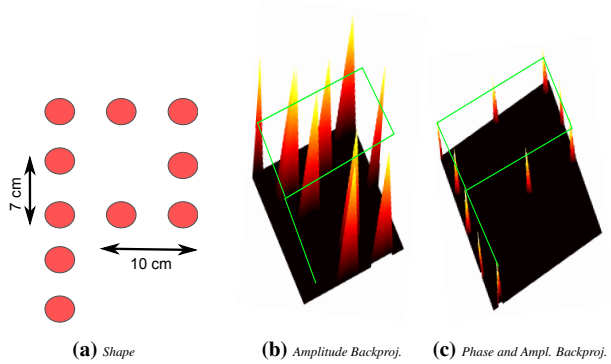


Figure 6: *Importance of phase measurements. (a) The point emitter is moved in the pattern of the letter "P" with precise spacing between points. (b) Result from amplitude backprojection (c) Result from backprojection using both phase and amplitude data. This problem is well referenced in the signal processing community.*

6 Do we need a TOF Camera?

Can the VSA model be applied to standard, digital camera measurements to reconstruct the occluded object? Objects further away appear to reflect less light due to the inverse square law. A reasonable question is whether amplitude information alone is sufficient to reconstruct the occluded object. In theory, the answer is yes, however, using amplitude information alone leads to an ill-conditioned inverse problem in the presence of noise. This is a case study which has already been well characterized within the source localization community, beginning with experiments in acoustic waves [Wallach 1939]. In fact, this is part of the reason why phased array systems—that measure boht phase and amplitude—are predominant in radar, sonar, and ultrasound imaging.

We empirically validate this to our context. As illustrated in Figure 6a, we move an active 30 MHz LED in the shape of the letter "P" and consider two approaches for reconstruction. Figure 6b illustrates the result when backprojecting the data using amplitude measurements alone. Note the result is unstable and does not match the geometry of the "P". In contrast Figure 6c solves the inverse problem using both phase and amplitude measurements to correctly backproject the location of the source.

Reproducibility

All computation in the paper (except superresolution with sparse priors) require only simple, computer vision calibration techniques and native commands from MATLAB. Since the hardware is also off-the-shelf, we believe this paper is easily reproduced by others in the field. Source code is also provided for the algorithm. In particular, running the source code will allow the user to reproduce Figures 8 and 9 from the main paper.

References

CHEN, S. S., DONOHO, D. L., AND SAUNDERS, M. A. 1998. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing 20*, 1, 33–61.

GOWER, J. C. 1985. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications 67*, 81–97.

NEEDELL, D., AND TROPP, J. A. 2009. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis 26*, 3, 301–321.

WALLACH, H. 1939. On sound localization. *The Journal of the Acoustical Society of America 10*, 4, 270–274.

WARD, G. 1992. Measuring and modeling anisotropic reflection. *Computer Graphics 26*, 2, 265–272.