# Optimal Structured Light à la Carte

Parsa Mirdehghan Wenzheng Chen Kiriakos N. Kutulakos
Department of Computer Science
University of Toronto
{parsa,wenzheng,kyros}@cs.toronto.edu

## Abstract

*We consider the problem of automatically generating sequences of structured-light patterns for active stereo triangulation of a static scene. Unlike existing approaches that use predetermined patterns and reconstruction algorithms tied to them, we generate patterns on the fly in response to generic specifications: number of patterns, projector-camera arrangement, workspace constraints, spatial frequency content,* etc. *Our pattern sequences are specifically optimized to minimize the expected rate of correspondence errors under those specifications for an unknown scene, and are coupled to a sequence-independent algorithm for per-pixel disparity estimation. To achieve this, we derive an objective function that is easy to optimize and follows from first principles within a maximum-likelihood framework. By minimizing it, we demonstrate automatic discovery of pattern sequences, in under three minutes on a laptop, that can outperform state-of-the-art triangulation techniques.*

## 1. Introduction

A key tenet in structured-light triangulation is that the choice of projection patterns matters a lot. Over the years, the field has seen significant boosts in performance—in robustness, 3D accuracy, speed and versatility—due to new types of projection patterns, and new vision algorithms tailored to them [1, 2]. These advances continue to this day, for improved robustness to indirect light [3–10]; computational efficiency [4, 11, 12]; high-speed imaging [13], outdoor 3D scanning [14, 15]; and for 3D imaging with specialized computational cameras [16], consumer-oriented devices [17, 18] and time-of-flight cameras [19–22].

Underlying all this work is a fundamental question: what are the optimal patterns to use and what algorithm should process the images they create? This question was originally posed by Horn and Kiryati twenty years ago [23] but the answer was deemed intractable and not pursued. Since then, pattern design has largely been driven by practical considerations [24–28] and by intuitive concepts borrowed from many fields (*e.g.*, communications [29], coding theory [2], number theory [25], numerical analysis [30], *etc.*)

In this paper we present the first computational approach to optimal design of patterns for structured light. We focus on the oldest, most accuracy-oriented version of this task: projecting a sequence of patterns one by one onto a static scene and using a camera to estimate per-pixel depth by triangulation. Starting from first principles, we formally derive an objective function over the space of pattern sequences that quantifies the expected number of incorrect stereo correspondences, and then minimize it using standard tools [31].

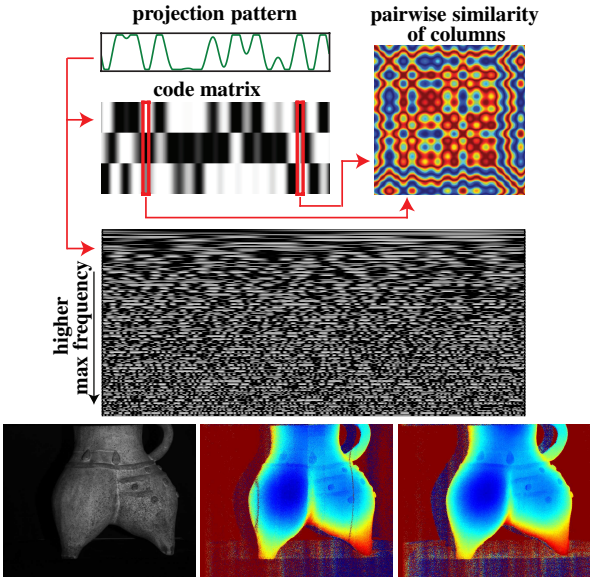


Figure 1: **Overview.** *Top:* A projection pattern is a 1D image projected along a projector's rows. A sequence of them defines a code matrix, whose columns encode pixel position. We present a framework for computing stereo correspondences using *optimal code matrices*, which we generate on the fly. These matrices minimize the expected number of stereo errors that occur when the individual matrix columns are not very distinctive (red=similar, blue=dissimilar). *Middle:* A whole space of optimal matrices exists, for different numbers of projection patterns, image signal-to-noise ratio, spatial frequency content (sample patterns shown above), *etc*. *Bottom:* We use two automatically-generated four-pattern sequences to compute the depth map of the object shown on left. Both are optimized for a one-pixel tolerance for stereo errors, without (middle) and with (right) a bounding-box constraint. Both depth maps are unprocessed (please zoom in).

Our optimization takes as input the projector's resolution and the desired number of projection patterns. In addition to these parameters, however, it can generate patterns that are precisely optimized for 3D accuracy on the system at hand (Figure 1): for the specific arrangement of projector

and camera; the shape and dimensions of the 3D scanning volume; the noise properties and peak signal-to-noise ratio of the overall imaging system; the defocus properties of the projector lens; a desired upper bound on the patterns' spatial frequency; and any unknown scene geometry. Thus, in contrast to prior work, we do not provide a closed-form expression or "codebook" for a one-size-fits-all pattern sequence; we give a way to generate scene-independent pattern sequences on the fly at near-interactive rates—less than three minutes on a standard laptop—so that the patterns and the associated reconstruction algorithm can be easily and automatically adapted for best performance. We call this paradigm *structured light à la carte*.

At the heart of our approach lies an extremely simple maximum-likelihood decoding algorithm for computing stereo correspondences independently of projection pattern. This algorithm is not only competitive with state-of-the-art pattern-specific decoders (MPS [3], EPS [4], XOR [5]), but also makes the pattern optimization problem itself tractable: by giving us a way to quantify the expected errors a pattern sequence will cause, it leads to an objective function over sequences that can be optimized numerically.

From a conceptual point of view our work makes four important contributions over the state of the art. First and foremost, our optimization-based approach turns structured-light imaging from a problem of *algorithm design* (for creating patterns [1], unwrapping phases [4, 32–34], computing correspondences [24], handling projector defocus [7, 35]) into one of *problem specification* (how many patterns, what working volume, what imaging system, *etc*.). The rest is handled automatically by pattern sequence optimization and maximum-likelihood decoding. Second, we demonstrate discovery of pattern sequences that can outperform state-of-the-art encoding schemes on hard cases: low numbers of patterns, geometrically-complex scenes, low signal-to-noise ratios. These are especially important in settings where speed and low-power imaging are of the essence. Third, the emergence of imaging systems that confer robustness to indirect light without restrictions on frequency content [16, 36] is giving us newfound degrees of freedom for pattern optimization; this larger design space can be explored automatically with our approach. Fourth, our formulation gives rise to new families of pattern sequences with unique properties, including (1) sequences designed to recover approximate, rather than exact, correspondences and (2) sequences designed with information about free space and stereo geometry already built in. This encodes geometric scene constraints directly into the optical domain for added reliability—via the patterns themselves—rather than enforcing them by post-processing less reliable 3D data.

## 2. Optimal Structured Light

Fundamentally, structured-light triangulation requires addressing two basic questions: (1) what patterns to project onto the scene and (2) how to compute projector-camera stereo correspondences from the images captured. Specifying a "good" set of projection patterns can be thought of as solving a one-dimensional *position encoding* problem for pixels on an epipolar line. Conversely, computing the stereo correspondence of a camera pixel can be thought of as a *position decoding* problem. We begin by formulating both problems in a probabilistic framework.

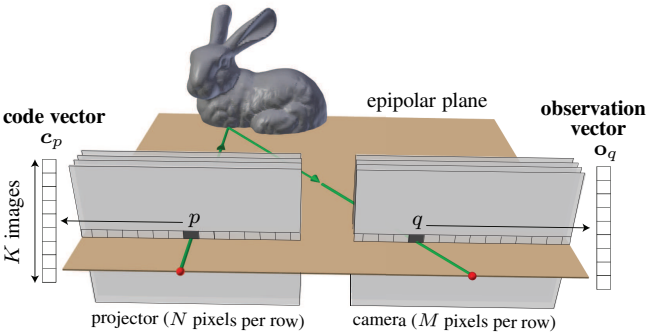

Figure 2: Viewing geometry. We assume the projector-camera system has been rectified, *i.e*., epipolar lines are along rows.

**The code matrix** A set of $K$ projection patterns implicitly assigns a $K$-dimensional *code vector* $\boldsymbol{c}_p$ to each pixel $p$ on the epipolar line (Figure 2). The elements of $\boldsymbol{c}_p$ are the pixel's intensity in the individual patterns, they can be non-binary, and must be chosen so that each code vector is as distinctive as possible. This becomes harder to do as $K$ decreases (*i.e*., vectors with fewer dimensions are less distinctive) and as the number of pixels increases (*i.e*., there are more vectors to be distinguished). We represent the code vectors of an epipolar line with a *code matrix* $\mathbf{C}$. This matrix has size $K \times N$ for an epipolar line with $N$ pixels.

**Position decoding** Consider a camera pixel $q$. The $K$ intensities observed at that pixel define a $K$-dimensional observation vector $\mathbf{o}_q$. Given this vector and the code matrix $\mathbf{C}$, the goal of position decoding is to infer its corresponding projector pixel $p^*$. This is a difficult problem because observations are corrupted by measurement noise and because the relation between observation vectors and code vectors can be highly non-trivial for general scenes. Following the communications literature [29], we formulate it as a maximum-likelihood (ML) problem:

$$p^* \;=\; \mathsf{Decode}(\mathbf{o}_q, \mathbf{C}) \tag{1}$$

$$\mathsf{Decode}(\mathbf{o}_q, \mathbf{C}) \;\stackrel{\text{def}}{=}\; \operatorname*{arg\,max}_{1 \le p \le N} \mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p) \;, \tag{2}$$

where $\mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p)$ is the likelihood that the code vector of pixel $q$'s true stereo correspondence is column $p$ of $\mathbf{C}$. This formulation is close in spirit to recent work on Bayesian time-of-flight depth estimation [22] but our image formation model and decoding procedure are very different. Note that the inferred correspondence $p^*$ may or may not agree with the true correspondence $p$ (Figure 2).

**Position encoding** The code matrix $\mathbf{C}$ should be chosen to minimize decoding error. For a given projector-camera sys-

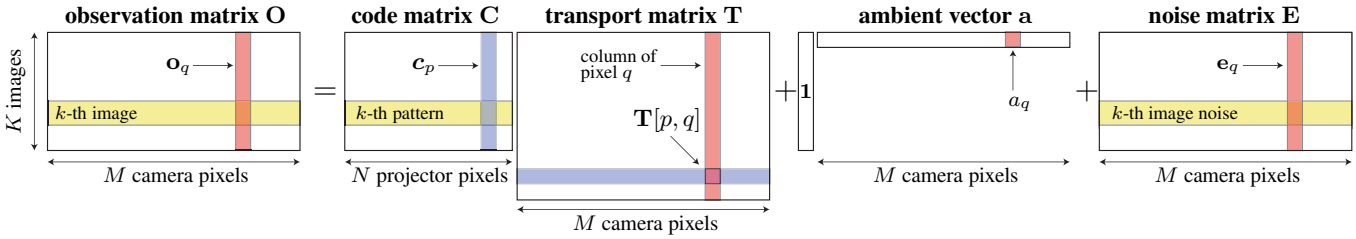


Figure 3: Generative model of image formation for a single epipolar line across $K$ images. Each column of matrix $\mathbf{O}$ is an observation vector (red) and each row collects the observations from a single image across all pixels on the epipolar line (yellow). All yellow rows are associated with the same input image and all red columns are associated with the same camera pixel $q$. The gray column and row are associated with the same projector pixel $p$.

tem and a specific scene, we quantify this error by counting the incorrect correspondences produced by the ML decoder:

$$\mathsf{Error}(\mathbf{C},\epsilon) \stackrel{\text{def}}{=} \sum_{q=1}^{M} \mathbb{1}\Big(\big|\mathsf{Decode}(\mathbf{o}_q,\mathbf{C}) - \mathsf{Match}(q)\big| > \epsilon\Big) \quad (3)$$

where $\mathsf{Match}(q)$ is the true stereo correspondence of image pixel $q$; $\epsilon$ is a tolerance threshold that permits small correspondence errors; $\mathbb{1}()$ is the indicator function; and the summation is over all pixels on the epipolar line. Note that evaluating the error function in Eq. (3) for a given scene and imaging system requires optimization, *i.e.*, solving the ML decoding problem in Eq. (2).

We now formulate optimal position encoding as the problem of finding a code matrix $\mathbf{C}^*_\epsilon$ that minimizes the expected number of incorrect correspondences:

$$\mathbf{C}^*_\epsilon \;=\; \arg\min_{\mathbf{C}} \mathbb{E}\big[\,\mathsf{Error}(\mathbf{C},\epsilon)\,\big] \quad (4)$$

where $\mathbb{E}\big[\,\big]$ denotes expectation over a user-specified domain of plausible scenes and imaging conditions. We call $\mathbf{C}^*_\epsilon$ the *optimal code matrix* for tolerance $\epsilon$.

**Key objective** We seek a solution to the nested optimization problem in Eq. (4) that is efficient to compute and can exploit imaging-system-specific information and user constraints. To do this, we cast the problem as an optimization in the space of *plausible epipolar transport matrices* (Section 3). This leads to a correlation-based ML decoder for structured-light reconstruction that is nearly optimal in low-noise settings (Section 4). Using this decoder, we derive a softmax-based approximation to the objective function of Eq. (4) and minimize it to get patterns that minimize the expected number of stereo mismatches (Section 5).

## 3. Epipolar-Only Image Formation

To simplify our formal analysis we assume that all light transport is *epipolar*. Specifically, we assume that observation vectors depend only on code vectors on the corresponding epipolar line. This condition applies to conventionally-acquired images when global light transport, projector defocus and camera defocus are negligible.[1] It also applies to all images captured by an epipolar-only imaging system regardless of scene content–even in the presence of severe global light transport [36].

When epipolar-only imaging holds and the system has been calibrated radiometrically, the relation between code vectors and observation vectors is given by (Figure 3):

$$\underbrace{\big[\mathbf{o}_1 \;\cdots\; \mathbf{o}_M\big]}_{\text{observation matrix } \mathbf{O}} = \underbrace{\big[\boldsymbol{c}_1 \;\cdots\; \boldsymbol{c}_N\big]}_{\text{code matrix } \mathbf{C}} \mathbf{T} + \mathbf{1}\,\underbrace{\big[a_1 \;\cdots\; a_M\big]}_{\text{ambient vector } \mathbf{a}} + \mathbf{E} \quad (5)$$

where $\mathbf{o}_1,\ldots,\mathbf{o}_M$ are the observation vectors of all pixels on an epipolar line; $a_1,\ldots,a_M$ are contributions of ambient illumination to these pixels; $\mathbf{1}$ is a column vector of all ones; matrix $\mathbf{E}$ is the observation noise; and $\mathbf{T}$ is the $N \times M$ epipolar transport matrix. Element $\mathbf{T}[p,q]$ of this matrix describes the total flux transported from projector pixel $p$ to camera pixel $q$ by direct surface reflection, global transport, and projector or camera defocus.

### 3.1. Plausible Epipolar Transport Matrices

The epipolar-only model of Eq. (5) encodes the geometry and reflectance of the scene as well as the scene's imaging conditions. It follows that the expectation in the position-encoding objective function of Eq. (4) is expressed most appropriately as an expectation over plausible epipolar transport matrices $\mathbf{T}$, ambient vectors $\mathbf{a}$, and noise matrices $\mathbf{E}$.

Let us first consider the space of plausible matrices $\mathbf{T}$. Even though the space of $N \times M$ matrices is extremely large, the matrices relevant to structured-light imaging belong to a much smaller space. This is because the elements of $\mathbf{T}$ associated with indirect light generally have far smaller magnitude than direct elements—and can thus be ignored [36]. This in turn makes likelihoods and expectations very efficient to compute. In particular, we consider ML-decoding and optimal encoding for the following three families:

**(A) Direct-only T, unconstrained:** The non-zero elements of $\mathbf{T}$ represent direct surface reflections and each camera pixel receives light from at most one projector pixel. It follows that each column of $\mathbf{T}$ contains at most one non-zero element. Moreover, the location of that element is a true stereo correspondence. The observation vector is therefore a noisy scaled-and-shifted code vector:

$$\mathbf{o}_q \;=\; \mathbf{T}[p,q]\cdot\boldsymbol{c}_p \;+\; a_q \;+\; \mathbf{e}_q \quad (6)$$

[1] See Figure 8 and [37] for experiments with scenes with significant indirect light, where this condition does not strictly hold.

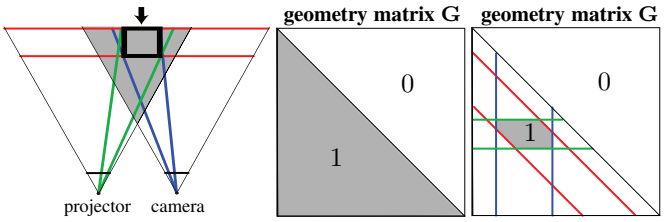


Figure 4: **Geometric constraints.** (a) Top view of the epipolar plane. (b) $\mathbf{T}$ is always lower triangular because the 3D rays of all other elements intersect behind the camera. (c) $\mathbf{T}$'s non-zero elements are restricted even further by knowledge of the working volume (*e.g.*, black square in (a)): its depth range (red) and its angular extent from the projector (green) and the camera (blue) define regions in $\mathbf{T}$ whose intersection contains all valid correspondences.

where vector $\mathbf{e}_q$ denotes noise. We assume that the location of the non-zero element in each column of $\mathbf{T}$ is drawn randomly from the set $\{1, \ldots, N\}$ and its value, $\mathbf{T}[p, q]$, is a uniform i.i.d random variable over $[0, 1]$. This amounts to being completely agnostic about the location and magnitude of $\mathbf{T}$'s non-zero elements.

**(B) Direct-only T with geometry constraints** We now restrict the above family to exclude geometrically-implausible stereo correspondences. These are elements of $\mathbf{T}$ whose associated 3D rays either intersect behind the image plane or outside a user-specified working volume (Figure 4a). We specify these invalid elements with a binary indicator matrix $\mathbf{G}$ (Figure 4b, 4c). Given this matrix, we assume that the location of the non-zero element in each column of $\mathbf{T}$ is drawn uniformly from the column's valid elements.

**(C) Direct-only T with projector defocus** The above two families do not model projector defocus. This not only prevents correct modeling of the defocused projection patterns that may illuminate some points [7], but also ignores the rich shape information available in the defocus cue [38]. Accounting for projector defocus in our framework is straightforward. Since a camera pixel may receive light from multiple projector pixels, the observation vector is a noisy scaled-and-shifted mixture of code vectors:

$$\mathbf{o}_q \;=\; \mathbf{T}[p, q] \cdot \left( \sum_{i=1}^{N} b_i^{pq} \boldsymbol{c}_i \right) \;+\; a_q \;+\; \mathbf{e}_q \qquad (7)$$

where $\mathbf{T}$ is a direct-only transport matrix from families (A) or (B). The coefficients $b_i^{pq}$ in Eq. (7) account for the defocus kernel. This kernel is depth dependent and thus each matrix element $\mathbf{T}[p, q]$ is associated with a different set of coefficients. The coefficients themselves can be computed by calibrating the projector [39]. Equation (7) can be made to conform to the epipolar image formation model of Eq. (5) by setting the scene's transport matrix to be a new matrix $\mathbf{T}'$ whose $i$-th row is $\mathbf{T}'[i, q] = \mathbf{T}[p, q] b_i^{pq}$. See [37] for details.

**Observation noise and ambient vector** The optimality of our ML position decoder (Section 4) relies on noise being signal independent and normally distributed. The position encoder of Section 5, on the other hand, can accommodate any model of sensor noise as long as its parameters are known. We assume that the elements of the ambient vector $\mathbf{a}$ follow a uniform distribution over $[0, a_{\max}]$, where $a_{\max}$ is the maximum contribution of ambient light expressed as a fraction of the maximum pixel intensity.

## 4. Optimal Position Decoding

Now suppose we are given a code matrix $\mathbf{C}$ and an observation vector $\mathbf{o}_q$ that conforms to the epipolar-only image formation model. Our task is to identify the stereo correspondence of pixel $q$. We seek a generic solution to this problem that does not impose constraints on the contents of the code matrix: it can contain code vectors defined *a priori*—such as MPS [3] or XOR [5] codes—or be a general matrix computed automatically through optimization.

Fortunately there is an extremely simple and near-optimal algorithm to do this: just compute the zero-mean normalized cross-correlation (ZNCC) [40] between $\mathbf{o}_q$ and the code vectors, and choose the one that maximizes it. This algorithm becomes optimal as noise goes to zero and as the variance of individual code vectors become the same:[2]

**Proposition 1 (ZNCC Decoding)** If observation vectors and code vectors are related according to Eq. (6) then

$$\lim_{\substack{v \to 0 \\ \sigma \to 0}} \left( \arg\max_{1 \le p \le N} \mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p) \right) \;=\; \arg\max_{1 \le p \le N} \mathsf{ZNCC}(\mathbf{o}_q, \boldsymbol{c}_p) \qquad (8)$$

where

$$\mathsf{ZNCC}(\mathbf{o}_q, \boldsymbol{c}_p) \;=\; \frac{\mathbf{o}_q - \mathsf{mean}(\mathbf{o}_q)}{\|\mathbf{o}_q - \mathsf{mean}(\mathbf{o}_q)\|} \cdot \frac{\boldsymbol{c}_p - \mathsf{mean}(\boldsymbol{c}_p)}{\|\boldsymbol{c}_p - \mathsf{mean}(\boldsymbol{c}_p)\|} \;, \qquad (9)$$

$v$ is the variance of the variances of the $N$ code vectors:

$$v \;=\; \mathsf{var}(\{\, \mathsf{var}(\boldsymbol{c}_1), \ldots, \mathsf{var}(\boldsymbol{c}_N)\}) \;, \qquad (10)$$

$\mathsf{mean}()$ and $\mathsf{var}()$ are over the elements of a code vector, $\sigma$ is the noise standard deviation, and $\mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p)$ is defined by marginalizing over ambient contributions and values of $\mathbf{T}[p, q]$:

$$\mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p) \stackrel{\text{def}}{=} \iint \mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p, \mathbf{T}[p, q] = x,\, a_q = y)\, \mathsf{Pr}(x)\, \mathsf{Pr}(x) dx dy \;.$$

**Definition 1 (ZNCC Decoder)**

$$\mathsf{Decode}(\mathbf{o}_q, \mathbf{C}) \;=\; \arg\max_{1 \le p \le N} \mathsf{ZNCC}(\mathbf{o}_q, \boldsymbol{c}_p) \;. \qquad (11)$$

**Corollary 1 (Defocused ZNCC Decoding)** If observation vectors and code vectors are related according to Eq. (7) then

$$\lim_{\substack{v \to 0 \\ \sigma \to 0}} \left( \arg\max_{1 \le p \le N} \mathsf{Pr}(\mathbf{o}_q \mid \boldsymbol{c}_p) \right) \;=\; \mathsf{Decode}(\mathbf{o}_q, \mathbf{C}\mathbf{T}^q) \qquad (12)$$

where the $N \times N$ matrix $\mathbf{T}^q$ holds the defocus kernel at camera pixel $q$ for all possible corresponding pixels $p$, *i.e.*, $\mathbf{T}^q[i, p] = b_i^{pq}$.

[2] See [37] for the proof. This result is a direct generalization of the widely known correlation-based ML decoder for communication channels corrupted by additive white Gaussian noise [29].

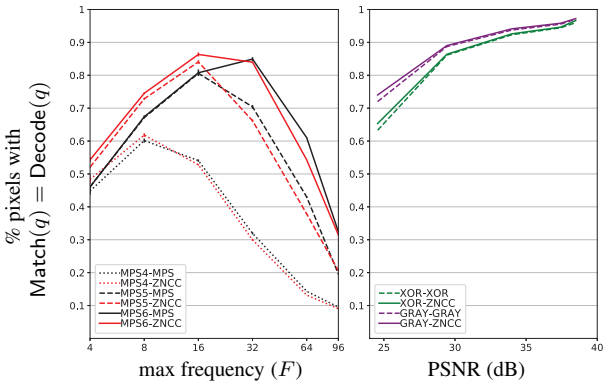


Figure 5: **ZNCC versus native decoding.** *Left:* We project $K$ MPS patterns [3] of maximum frequency $F$ onto a known planar target and compute correspondence errors using our ZNCC decoder (red) and the one by the MPS authors (black). *Right:* A similar comparison for 10 Gray codes (purple) and 10 XOR-04 codes (green), projected along with their binary complement. We used the binarization technique in [24] for "native" decoding. Since these codes have no frequency bound we plot them against image PSNR. In all cases, ZNCC decoding yields comparable results.

The near-optimality of the ZNCC decoder has two important implications. First, it suggests that there is potentially no accuracy advantage to be gained by designing decoding algorithms tailor-made for specific codes[3] (Figure 5). Second, it allows us to transform the nested position-encoding optimization of Eq. (4) into a conventional non-linear optimization. This opens the door to automatic generation of optimized code matrices, discussed next.

## 5. Optimal Position Encoding

We begin by developing a continuous approximation to the function $\mathsf{Error}()$ in Eq. (3). This function counts the decoding errors that occur when a given code matrix $\mathbf{C}$ is applied to a specific scene and imaging condition, *i.e.*, a specific transport matrix $\mathbf{T}$, observation noise $\mathbf{E}$, and ambient vector $\mathbf{a}$. To evaluate the position-encoding objective function on matrix $\mathbf{C}$, we draw $S$ fair samples over $\mathbf{T}$, $\mathbf{E}$ and $\mathbf{a}$:

$$\mathbb{E}\big[\,\mathsf{Error}(\mathbf{C},\epsilon)\,\big] = (1/S)\sum_{\mathbf{T},\mathbf{E},\mathbf{a}} \mathsf{Error}(\mathbf{T},\mathbf{E},\mathbf{a},\mathbf{C},\epsilon)\ . \tag{13}$$

**Softmax approximation of decoding errors** Consider a binary variable that tells us whether or not the optimal decoder matched camera pixel $q$ to a projector pixel $p$. We approximate this variable by a continuous function in three steps using Eqs. (15)-(17): Equation (15) states that in order for projector pixel $p$ to be matched to $q$, the likelihood of $p$'s code vector must be greater than all others. Equation (16) then follows from Proposition 1, allowing us to replace likelihoods with ZNCC scores. Last but not least, Eq. (17) approximates the indicator variable with a softmax ratio; as the scalar $\mu$ goes to infinity, the ratio tends to 1 if pixel $p$'s ZNCC score is the largest and tends to 0 otherwise:

$$\mathbb{1}\Big(\,|\,\mathsf{Decode}(\mathbf{o}_q,\mathbf{C})-p\,|=0\Big) \ = \tag{14}$$

$$= \ \mathbb{1}\Big(\mathsf{Pr}(\mathbf{o}_q\mid \boldsymbol{c}_p) = \max_{1\le r\le N}\mathsf{Pr}(\mathbf{o}_q\mid \boldsymbol{c}_r)\Big) \tag{15}$$

$$\underset{\substack{v\to 0\\ \sigma\to 0}}{=} \ \mathbb{1}\Big(\mathsf{ZNCC}(\mathbf{o}_q,\boldsymbol{c}_p) = \max_{1\le r\le N}\mathsf{ZNCC}(\mathbf{o}_q,\boldsymbol{c}_r)\Big) \tag{16}$$

$$\underset{\mu\to\infty}{=} \ \frac{\exp\Big(\mu\cdot\mathsf{ZNCC}(\mathbf{o}_q,\boldsymbol{c}_p)\Big)}{\sum_{r=1}^{N}\exp\Big(\mu\cdot\mathsf{ZNCC}(\mathbf{o}_q,\boldsymbol{c}_r)\Big)} \tag{17}$$

$$\stackrel{\text{def}}{=} \ f_\mu(\mathbf{C},\mathbf{o}_q,p)\ . \tag{18}$$

To count all correct matches on an epipolar line, we evaluate the softmax ratio at the true stereo match of every pixel $q$, and then compute their sum. Using the notation in Eq. (18):

$$\mathsf{Correct}(\mathbf{T},\mathbf{E},\mathbf{a},\mathbf{C}) \ = \ \sum_{q=1}^{M} f_\mu(\mathbf{C},\mathbf{o}_q,\mathsf{Match}(q))\ . \tag{19}$$

Finally, incorporating the tolerance parameter $\epsilon$ to permit small errors in stereo correspondences we get:

$$\mathsf{Correct}(\mathbf{T},\mathbf{E},\mathbf{a},\mathbf{C},\epsilon) = \sum_{q=1}^{M}\sum_{r=-\epsilon}^{\epsilon} f_\mu(\mathbf{C},\mathbf{o}_q,\mathsf{Match}(q)+r) \tag{20}$$

$$\mathsf{Error}(\mathbf{T},\mathbf{E},\mathbf{a},\mathbf{C},\epsilon) \ = \ M - \mathsf{Correct}(\mathbf{T},\mathbf{E},\mathbf{a},\mathbf{C},\epsilon)\ . \tag{21}$$

**Sampling scenes and imaging conditions** Constructing fair samples of the observation noise $\mathbf{E}$ and ambient vector $\mathbf{a}$ is straightforward and omitted. To construct a direct-only matrix whose geometric constraints are a matrix $\mathbf{G}$, we proceed as follows. We first randomly assign a valid stereo correspondence to each camera pixel according to $\mathbf{G}$. This specifies the location of the single non-zero element in each column of $\mathbf{T}$ (Figure 3). We then assign a random value to each of those elements independently. The result is a valid direct-only transport matrix, *i.e.*, a sample from family (B) in Section 4. To construct a family-(C) sample $\mathbf{T}'$ that accounts for projector defocus and geometric constraints, we construct a direct-only matrix $\mathbf{T}$ according to $\mathbf{G}$ and then incorporate the depth-dependent defocus kernels (Section 4).

**Optimization** We use the Adam optimizer [31] to perform stochastic gradient descent on the objective function in Eq. (13) with a fixed learning rate of 0.01. The user-specified parameters are (1) the number of projector pixels $N$; (2) the number of camera pixels $M$; (3) the number of projection patterns $K$; (4) the desired tolerance parameter $\epsilon$; and (5) the geometric constraint matrix $\mathbf{G}$. The result of the optimization is a code matrix $\mathbf{C}^*_\epsilon$.

We initialize the optimization with a random $K\times N$ code matrix $\mathbf{C}$ and draw a total of $S=500$ samples $(\mathbf{T},\mathbf{E},\mathbf{a})$ at iteration 1 to define the objective function of Eq. (13).

[3] Strictly speaking this applies to decoders that estimate depth at each pixel independently, and to code vectors that have approximately the same variance. Note that ZNCC decoding does incur a computational penalty: it requires $O(N)$ operations versus $O(K)$ of most specialized decoders.

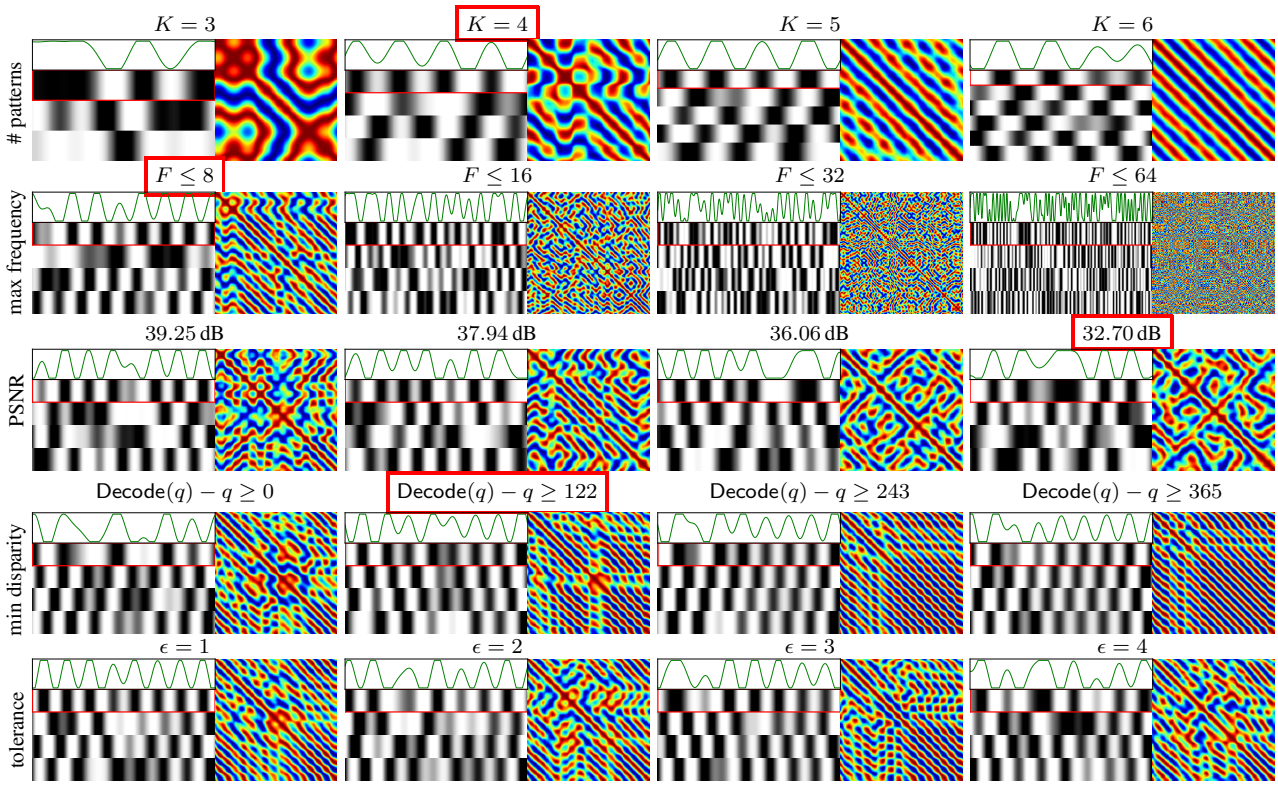


Figure 6: **A walk in the space of optimal codes.** To better visualize code structure, the pairwise scores $\mathsf{ZNCC}(\boldsymbol{c}_i, \boldsymbol{c}_j)$ of code vectors are shown as a jet-color-mapped matrix (deep red $= 1$, deep blue $= -1$). These can be treated as a confusion matrix. *Row 1:* We set the maximum spatial frequency of the patterns to $F = 4$ and the image PSNR to be maximal for our imaging conditions (frame rate=$50Hz$, camera gain=1, known read noise, pixel intensity that spans the full interval $[0, 1]$). We then compute the optimal code matrix for our 608-pixel projector for different numbers of patterns and no other constraints. *Row 2:* We then choose $K = 4$ (outlined in red in Row 1) and compute optimal matrices for different bounds on the maximum spatial frequency, with everything else fixed as above. *Row 3:* We now set the frequency to 8 (outlined in red in Row 2) and compute optimal matrices for different values of pixel PSNR (*i.e.*, the maximum image intensity gets increasingly smaller), again with everything else fixed as above. *Rows 4 and 5:* We follow the exact same process for different lower bounds on disparity (*i.e.*, the maximum scene depth is increasingly being restricted), and different tolerances in correspondence error.

These samples act as a "validation set" and remain fixed until convergence. For gradient calculations we use a mini-batch containing two new randomly-drawn samples per iteration. Optimization converges in around 250 iterations (152 seconds on an 8-core 2.3GHz MacBook Pro laptop for a six-pattern matrix). We found that increasing the number of samples had no appreciable effect on the quality of $\mathbf{C}^*_\epsilon$ (*i.e.*, the number of decoding errors on other randomly-generated scenes and imaging conditions). What does make a big difference, however, is the value of the softmax multiplier $\mu$: there is significant degradation in quality for $\mu < 300$, but increasing it beyond that value has little effect. We use $\mu = 300$ for all results shown. See [37] for more details.

**Frequency-constrained projection patterns** Many structured-light techniques advocate use of projection patterns with spatial frequency no larger than a user-specified threshold $F$ [3, 4, 41]. This can be viewed as an additional design constraint on the optimal code matrix. To explicitly enforce it, we project the code matrix computed at each iteration onto the space of matrices satisfying the constraint.[4]

[4]This can be achieved using a projection-onto-convex-set (POCS) algorithm [42]; in practice, two iterations will yield a close point to the feasible space. Specifically, for each row of $\mathbf{C}$, we (1) clip its elements to $[0, 1]$, (2) compute the Discrete Fourier Transform, (3) set to zero all DFT coefficients above $F$, (4) compute the inverse DFT of the result, and (5) clip it to $[0, 1]$ again. This yields a code matrix $\mathbf{C}'$ that is fed to the next iteration.

**Advanced sensor noise modeling** Although the ZNCC decoder is optimal only for additive Gaussian noise, the objective function in Eq.(13) can incorporate any sensor noise model: we simply draw samples of $\mathbf{E}$ from the camera's noise distribution. We found that this improves significantly the real-world performance of the optimized codes.

## 6. Experimental Results

**The space of optimal code matrices** Figure 6 shows several code matrices generated by our optimizer. It is clear by inspection that the codes exhibit a very diverse structure that adapts significantly in response to user specifications. Increasing the frequency content (Row 2) produces confusion matrices with much less structure—as one would intuitively expect from vectors that are more

distinctive. Interestingly, codes adapted to lower peak signal-to-noise ratio (PSNR) conditions have confusion matrices with coarser structure. We did not, however, observe an appreciable difference in the real-world performance of those matrices. Row 3 of Figure 6 illustrates the codes' adaptation to geometric constraints. Specifically, only points on the plane at infinity can have $\mathsf{Decode}(q)=q$ and for 3D points that are closer, a camera pixel can only be matched to a projector pixel on its right (Figure 4b). Comparing the code matrix for an unrestricted $\mathbf{T}$ (red box on Row 3) to that of a lower-triangular $\mathbf{T}$ (first column in Row 4) one sees significant re-organization in the confusion matrix; the optimization effectively "focuses" the codes' discriminability to only those code vectors that yield valid 3D points. On the other hand, code matrices that compute approximate, rather than exact correspondences, exhibit coarser structure in their confusion matrix (Row 4).

**Experimental system** We acquired all images at $50Hz$ and $8$ bits with a $1280\times1024$ monochrome camera supplied by IDS (model IDS UI-3240CP-M), fitted with a Lensation F/1.6 lens (model CVM0411). For pattern projection we used a 100-lumen DLP projector by Keynote Photonics (model LC3000) with a native resolution of $608 \times 684$ and only the red LED turned on. We disabled gamma correction, verified the system's linear radiometric response, and measured the sensor's photon transfer curve. This made it possible to get a precise measure of PSNR independently for each pixel on the target. We experimented with three different models of pixel noise for our position-encoding optimization: (1) additive Gaussian, (2) Poisson shot noise with additive read noise, and (3) exponential noise [43] with additive read noise.

**Ground truth** We printed a random noise pattern of bounded frequency onto a white sheet of paper and placed it on a planar target $60$cm away from the stereo pair (Figure 7, bottom row, third column). We used two different pattern sequences to obtain "ground-truth" disparity maps: $160$ conventional phase-shifted patterns and $20$ XOR patterns (including the complement codes). We adjusted the aperture so that the maximum image intensity was 200 for a white projection pattern (*i.e.*, a high-PSNR regime at the brightest pixels) and focused the lens on the target. For $97\%$ of pixels the disparities were identical in the two maps; the rest differed by $\pm1$ disparity. Thus, correctness above $97\%$ against these maps is not significant. We optimize all of our code matrices for these high-PSNR conditions with the exponential-plus-read-noise model (see [37] for performance evaluation of other two models).

**Quantitative evaluation** We focus here on the most challenging cases: very small number of patterns and low PSNR. To evaluate low-PSNR performance, we reduced the aperture so that the brightest pixel intensity under a white projection pattern is $60$, and counted the pixels whose correspondences are within $\epsilon$ of the ground truth. Figure 7 compares our optimized code matrices against those of MPS and EPS, using the same ZNCC decoder for all codes. Several observations can be made from these results. First, our code matrices outperform MPS and EPS—which represent the current state of the art—in all cases shown. This performance gap, however, shrinks for larger numbers of patterns. Second, our codes perform significantly better than EPS and MPS at higher spatial frequencies. This is despite the fact that those coding schemes were specifically designed to produce high-frequency patterns. It is also worth noting that the performance degradation of MPS and EPS at high frequencies cannot be explained by camera defocus because the camera's aperture was small in these experiments (*i.e.*, large depth of field). Third, geometric constraints confer a major performance advantage to all codes at low pattern counts. The gain, however, is higher for our codes since they are optimized precisely for them. Fourth, code matrices that are geometry-constrained and optimized for a small error tolerance tend to produce low root-mean-squared errors (RMSE) for most frequencies.

**Qualitative results** Reconstructions of several objects are shown in Figure 1 (using four patterns) and Figure 8 (using five and six patterns). The comparison in Figure 1 indicates that computing geometry-constrained codes has a clear effect on the quality of the results—a trend observed in our quantitative comparisons as well. In Figure 8 we specifically chose to reconstruct a dark scene as well as a scene with significant indirect light to compare performance under low-PSNR conditions and general light transport. We observe that our depth maps have significantly fewer outliers than EPS and MPS and are less influenced by depth discontinuities. Moreover, despite not being specifically optimized for indirect light, we obtain better depth maps there as well.

## 7. Concluding Remarks

We believe this is just a small first step in designing optimized codes for structured light. The specific imaging regime we chose to study—small numbers of patterns, low-PSNR conditions—leaves a lot of room for further exploration. On the high-accuracy end of the spectrum, the ability to quickly optimize patterns after an initial 3D scan could lead the way to new adaptive 3D scanning techniques [5, 44].

Although derived from first principles, our position-encoding objective function can be viewed as an extremely simple one-layer neural network. Understanding how to best exploit the power of deeper architectures for active triangulation is an exciting direction for future work.

**Acknowledgements** The support of the Natural Sciences and Engineering Council of Canada under the RGPIN and SGP programs, and of DARPA under the REVEAL program, are gratefully acknowledged.

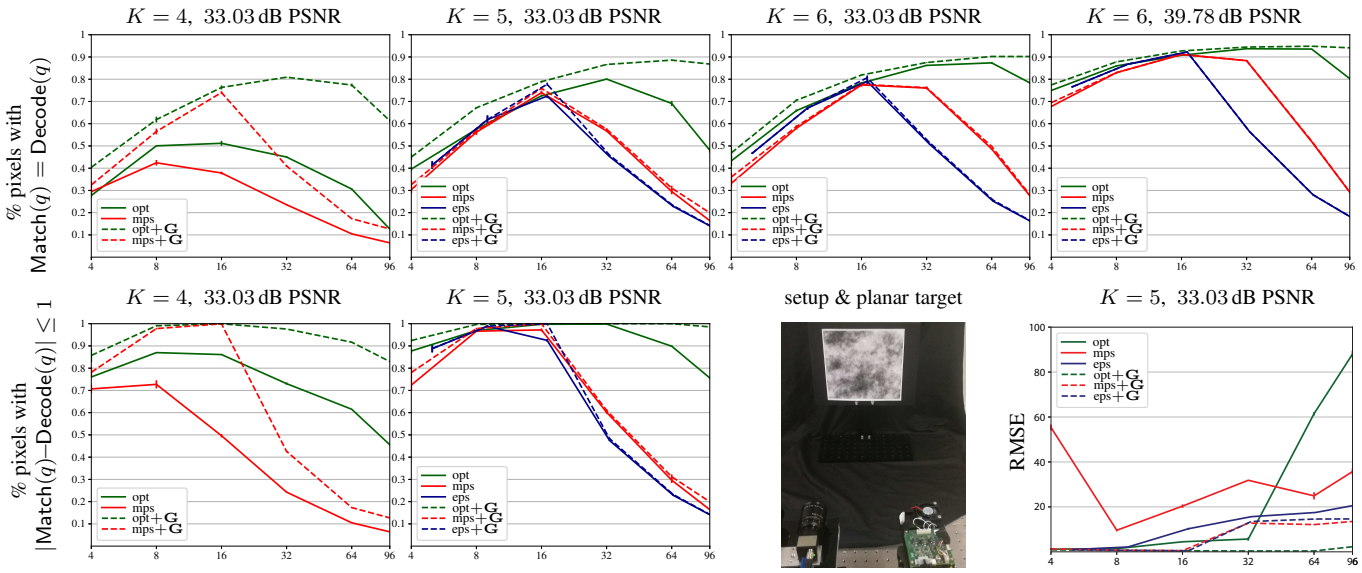


Figure 7: **Quantitative evaluation.** *Top row and first two columns of bottom row:* Each data point represents three independent acquisitions with the same pattern sequence. Error bars indicate the smallest and largest fraction of correct correspondences in those runs. We used $\epsilon = 0$ for optimization in the top row and $\epsilon = 1$ in the bottom. Solid lines show results when no geometry constraints are imposed on code optimization and on decoding. Dashed lines show what happens when we use a depth-constrained geometry matrix $\mathbf{G}$ (Figure 4c). For EPS and MPS, the constraint is used only for decoding, *i.e.*, we search among the valid correspondences for the one that maximizes the ZNCC score. Our codes, on the other hand, are optimized for that constraint and decoded with it as well. *Bottom row, right: RMSE plots.*

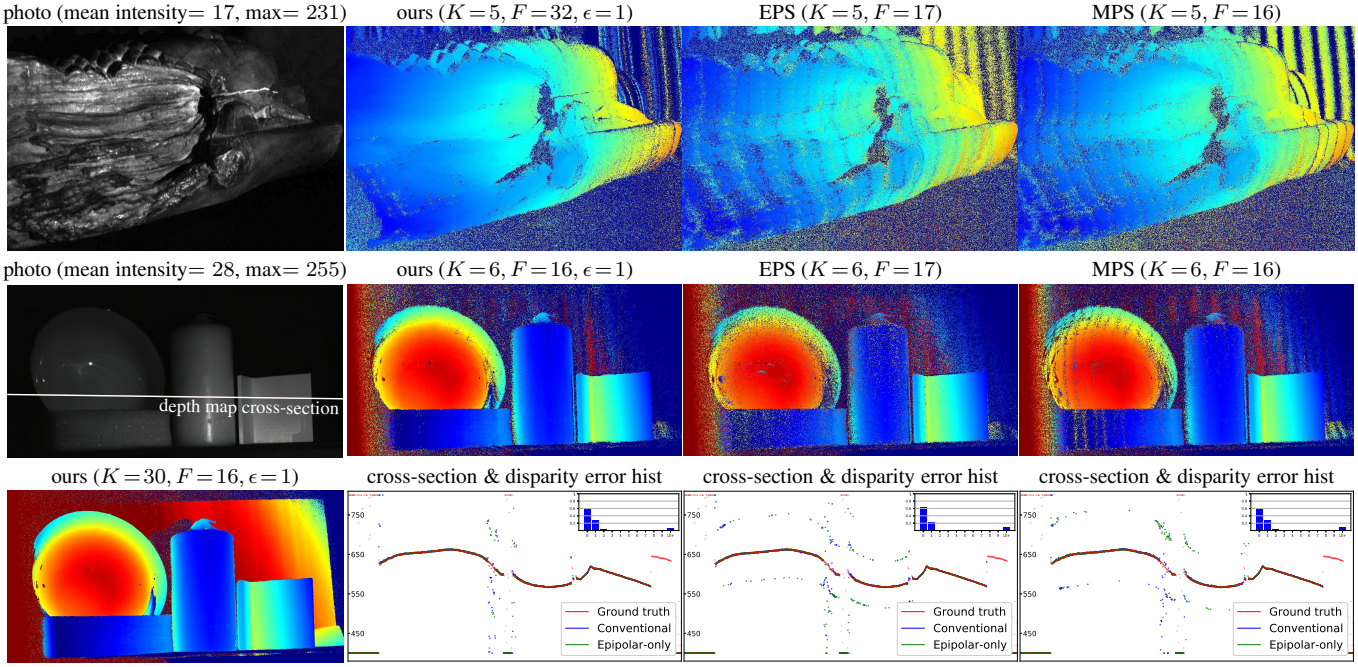


Figure 8: **Qualitative comparisons.** We acquired depth maps for the scenes on the left using three methods, with the same ZNCC decoder and the same triangular geometry matrix $\mathbf{G}$ (Figure 4b). For each method, we reconstructed the scenes for several maximum frequencies in the range $[4, 32]$ and show depth maps for each method's best-performing frequency. *Top row:* Reconstructing a dark, varnished and sculpted wooden trunk with five patterns. *Middle row:* Reconstructing a scene with significant indirect transport (a bowl, candle, and convex wedge) using conventional imaging and six patterns. *Bottom row:* Depth map acquired with many more patterns, along with cross-sections of the above depth maps (blue points) and a histogram of disparity errors (please zoom in to the electronic copy). For reference, we include the cross-sections of depth maps acquired using epipolar-only imaging [36] with the exact same patterns (green points), as well as of "ground truth" depth maps acquired with 160 shifted cosine patterns of frequencies 16 to 31 using epipolar-only imaging (red points).

## References


[1] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern Recogn*, vol. 43, no. 8, pp. 2666–2680, 2010.

[2] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recogn*, vol. 37, no. 4, pp. 827–849, 2004.

[3] M. Gupta and S. Nayar, "Micro Phase Shifting," in *Proc. IEEE CVPR*, pp. 813–820, 2012.

[4] D. Moreno, K. Son, and G. Taubin, "Embedded phase shifting: Robust phase shifting with embedded signals," in *Proc. IEEE CVPR*, pp. 2301–2309, 2015.

[5] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "A Practical Approach to 3D Scanning in the Presence of Interreflections, Subsurface Scattering and Defocus," *Int. J. Computer Vision*, vol. 102, no. 1-3, pp. 33–55, 2013.

[6] J. Gu, T. Kobayashi, M. Gupta, and S. K. Nayar, "Multiplexed illumination for scene recovery in the presence of global illumination," in *Proc. IEEE ICCV*, pp. 691–698, 2011.

[7] M. Gupta, Y. Tian, S. G. Narasimhan, and L. Zhang, "A Combined Theory of Defocused Illumination and Global Light Transport," *Int. J. Computer Vision*, vol. 98, no. 2, 2011.

[8] V. Couture, N. Martin, and S. Roy, "Unstructured Light Scanning Robust to Indirect Illumination and Depth Discontinuities," *Int. J. Computer Vision*, vol. 108, no. 3, pp. 204–221, 2014.

[9] T. Chen, H.-P. Seidel, and H. P. A. Lensch, "Modulated phase-shifting for 3D scanning," in *Proc. IEEE CVPR*, 2008.

[10] Y. Xu and D. G. Aliaga, "Robust pixel classification for 3d modeling with structured light," in *Proc. Graphics Interface*, pp. 233–240, 2007.

[11] S. R. Fanello, C. Rhemann, V. Tankovich, A. Kowdle, S. O. Escolano, D. Kim, and S. Izadi, "HyperDepth: Learning Depth from Structured Light without Matching," in *Proc. IEEE CVPR*, pp. 5441–5450, 2016.

[12] S. R. Fanello, J. Valentin, C. Rhemann, A. Kowdle, V. Tankovich, P. Davidson, and S. Izadi, "UltraStereo: Efficient Learning-Based Matching for Active Stereo Systems," in *Proc. IEEE CVPR*, pp. 6535–6544, 2017.

[13] S. J. Koppal, S. Yamazaki, and S. G. Narasimhan, "Exploiting DLP Illumination Dithering for Reconstruction and Photography of High-Speed Scenes," *Int. J. Computer Vision*, vol. 96, no. 1, pp. 125–144, 2012.

[14] M. Gupta, Q. Yin, and S. Nayar, "Structured Light in Sunlight," in *Proc. IEEE ICCV*, pp. 545–552, 2013.

[15] C. Mertz, S. J. Koppal, S. Sia, and S. G. Narasimhan, "A low-power structured light sensor for outdoor scene reconstruction and dominant material identification," in *IEEE PROCAMS*, pp. 15–22, 2012.

[16] M. O'Toole, S. Achar, S. G. Narasimhan, and K. N. Kutulakos, "Homogeneous codes for energy-efficient illumination and imaging," in *Proc. ACM SIGGRAPH Asia*, 2015.

[17] D. Moreno, F. Calakli, and G. Taubin, "Unsynchronized structured light," in *Proc. ACM SIGGRAPH Asia*, pp. 178–11, 2015.

[18] M. Donlic, T. Petkovic, and T. Pribanic, "On Tablet 3D Structured Light Reconstruction and Registration," in *Proc. ICCV*, pp. 2462–2471, 2017.

[19] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar, "Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles," in *Proc. ACM SIGGRAPH Asia*, 2013.

[20] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar, "Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization," *Optics Letters*, vol. 39, no. 6, pp. 1705–1708, 2014.

[21] G. Satat, M. Tancik, and R. Raskar, "Lensless imaging with compressive ultrafast sensing," *IEEE Trans. Comput. Imaging*, vol. 3, no. 3, pp. 398–407, 2017.

[22] A. Adam, C. Dann, O. Yair, S. Mazor, and S. Nowozin, "Bayesian Time-of-Flight for Realtime Shape, Illumination and Albedo," *IEEE T-PAMI*, vol. 39, no. 5, pp. 851–864, 2017.

[23] E. Horn and N. Kiryati, "Toward optimal structured light patterns," in *Proc. IEEE 3DIM*, pp. 28–35, 1997.

[24] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE CVPR*, pp. 195–202, 2003.

[25] T. Pribanic, H. Džapo, and J. Salvi, "Efficient and Low-Cost 3D Structured Light System Based on a Modified Number-Theoretic Approach," *EURASIP J. Adv. Signal Process.*, 2010.

[26] T. Pribanic, S. Mrvoš, and J. Salvi, "Efficient multiple phase shift patterns for dense 3D acquisition in structured light scanning," *Image and Vision Computing*, vol. 28, no. 8, pp. 1255–1266, 2010.

[27] L. Zhang, B. Curless, and S. M. Seitz, "Rapid shape acquisition using color structured light and multi-pass dynamic programming," in *Proc. IEEE 3DPVT*, pp. 24–36, 2002.

[28] M. Holroyd, J. Lawrence, and T. Zickler, "A Coaxial Optical Scanner for Synchronous Acquisition of 3D Geometry and Surface Reflectance," in *Proc. ACM SIGGRAPH Asia*, 2010.

[29] J. G. Proakis, *Digital Communications*. McGraw-Hill, 2001.

[30] P. Leopardi, *Distributing points on the sphere*. PhD thesis, University of New South Wales, Apr. 2007.

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.

[32] J. M. Huntley and H. Saldner, "Temporal phase-unwrapping algorithm for automated interferogram analysis," *Appl Optics*, vol. 32, no. 17, pp. 3047–3052, 1993.

[33] R. Ishiyama, S. Sakamoto, J. Tajima, T. Okatani, and K. Deguchi, "Absolute phase measurements using geometric constraints between multiple cameras and projectors," *Appl Optics*, vol. 46, no. 17, pp. 3528–3538, 2007.

[34] C. Peters, J. Klein, M. B. Hullin, and R. Klein, "Solving trigonometric moment problems for fast transient imaging," *ACM Trans. Graphics*, vol. 34, no. 6, p. 220, 2015.

[35] S. Achar and S. G. Narasimhan, "Multi Focus Structured Light for Recovering Scene Shape and Global Illumination," in *Proc. ECCV*, pp. 205–219, 2014.

[36] M. O'Toole, J. Mather, and K. N. Kutulakos, "3D Shape and Indirect Appearance by Structured Light Transport," *IEEE T-PAMI*, vol. 38, no. 7, pp. 1298–1312, 2016.

[37] P. Mirdehghan, W. Chen, and K. N. Kutulakos, "Optimal Structured Light a la Carte: Supplemental Document," in *Proc. IEEE CVPR*, 2018. Also available at http://www.dgp.toronto.edu/OptimalSL.

[38] S. K. Nayar, M. Watanabe, and M. Noguchi, "Real-time focus range sensor," *IEEE T-PAMI*, vol. 18, no. 12, pp. 1186–1198, 1996.

[39] L. Zhang and S. K. Nayar, "Projection defocus analysis for scene capture and image display," in *Proc. ACM SIGGRAPH*, pp. 907–915, 2006.

[40] J. Martin and J. L. Crowley, "Experimental Comparison of Correlation Techniques ," in *Int. Conf. on Intelligent Autonomous Systems*, 1995.

[41] S. K. Nayar, G. Krishnan, M. D. Grossberg, and R. Raskar, "Fast separation of direct and global components of a scene using high frequency illumination," in *Proc. ACM SIGGRAPH*, pp. 935–944, 2006.

[42] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM review*, vol. 38, no. 3, pp. 367–426, 1996.

[43] F. J. Spang, I. Rosenberg, E. Hedin, and G. Royle, "Photon small-field measurements with a CMOS active pixel sensor," *Phys. Med. Biol.*, vol. 60, pp. 4383–4398, May 2015.

[44] T. Koninckx and L. Van Gool, "Real-time range acquisition by adaptive structured light," *IEEE T-PAMI*, vol. 28, no. 3, pp. 432–445, 2006.