3D Shape and Indirect Appearance By Structured Light Transport

Matthew O'Toole John Mather Kiriakos N. Kutulakos
Department of Computer Science
University of Toronto
{motoole,jmather,kyros}@cs.toronto.edu

Abstract

We consider the problem of deliberately manipulating the direct and indirect light flowing through a time-varying, fully-general scene in order to simplify its visual analysis. Our approach rests on a crucial link between stereo geometry and light transport: while direct light always obeys the epipolar geometry of a projector-camera pair, indirect light overwhelmingly does not. We show that it is possible to turn this observation into an imaging method that analyzes light transport in real time in the optical domain, prior to acquisition. This yields three key abilities that we demonstrate in an experimental camera prototype: (1) producing a live indirect-only video stream for any scene, regardless of geometric or photometric complexity; (2) capturing images that make existing structured-light shape recovery algorithms robust to indirect transport; and (3) turning them into one-shot methods for dynamic 3D shape capture.

1. Introduction

A common assumption in computer vision is that light travels along *direct* paths, *i.e.*, it goes from source to camera by bouncing at most once in the scene. While this assumption works well in many cases, light propagation through natural scenes is actually a much more complex phenomenon: light reflects and refracts, it undergoes specular and diffuse inter-reflections, it scatters volumetrically and creates caustics, and may do all of the above in the same scene. Analyzing all these phenomena with a conventional camera is a hard, open problem—and is even harder when the scene is dynamic and light transport changes unpredictably.

Despite the problem's intrinsic difficulty, indirect transport is a major component of real-world appearance [1] and an important cue for scene and material understanding [2]. It is also a major factor preventing broader use of structured-light techniques, which largely assume direct or low-frequency light transport (*e.g.*, 3D laser scanning [3, 4], active triangulation [5, 6] and photometric stereo [7]).

As a step toward analyzing scenes that exhibit complex light transport, in this paper we develop a framework for imaging them in real time. Our focus is on the general case where the scene is unknown; its motion and photometric properties unrestricted; and its illumination comes from one or more controllable sources in general position (*e.g.*, projectors).

Working from first principles, we show that two families of transport paths dominate image formation in a projector-camera system: *epipolar paths*, which satisfy the familiar epipolar constraint and contribute to a scene's direct image, and *non-epipolar paths* which contribute to its indirect. Crucially, while the contributions of these paths are hard to separate in an image, the paths themselves are easy to untangle in the optical domain *before* acquisition takes place. Using this idea as a starting point, we develop a novel technique called *Structured Light Transport (SLT)* that processes epipolar and non-epipolar paths optically for the purpose of live imaging and 3D shape recovery. In particular, we define and address four imaging problems:

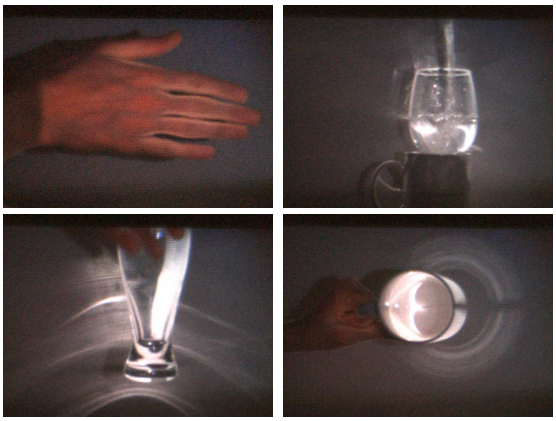

Figure 1: Snapshots from raw live indirect video. Clockwise from top: (1) A hand; note the vein pattern and the inter-reflections between fingers. (2) Pouring water into a glass. (3) Caustics formed inside a mug from specular inter-reflections; note the secondary reflections to the board behind the mug and from the board onto the mug's exterior surface. (4) Refractions and caustics from a beer glass. See Figure 9 for more images and [8] for videos.

- **one-shot indirect-only imaging:** capture an image that records only contributions from indirect light;
- **one-shot indirect-invariant imaging:** given any desired illumination, capture an image where light appears to have been transported by direct paths only;
- **two-shot direct-only imaging:** capture two images whose difference contains only the direct light; and
- **one-shot multi-pattern imaging:** given any $N \geq 2$ desired illuminations, capture an image that "packs" into one shot N separate views of the scene, each corresponding to a desired illumination.

Little is currently known about how to solve these prob-

lems in the general setting we consider. Our solutions, while firmly rooted in computer vision, operate exclusively in the optical domain and require no computational post-processing: our implementation is a physical device that just outputs live video; this is optionally processed "downstream" by standard 3D reconstruction algorithms [5] which can be oblivious to the complexity of light transport occurring in a scene. The device itself is a novel combination of existing off-the-shelf components—a conventional video camera operating at 28Hz, a pair of synchronized digital micro-mirror devices (DMDs) operating at 2.7kHz to 24kHz, and optics for coupling them.

From a practical point of view, our work offers four main contributions over the state of the art. First, it is the first demonstration of an "indirect-only video camera," *i.e.*, a camera that outputs a live stream of indirect-only video for fully-general scenes—exhibiting arbitrary motion, caustics, specular inter-reflections and numerous other transport effects. Prior work on indirect imaging was either constrained to static scenes [9, 10], or assumed diffuse/low-frequency transport [2, 11] and accurate 2D motion estimation [11]. Second, we show how to capture—with just one SLT shot—views of a scene that are invariant to indirect light. This is particularly useful for imaging dynamic scenes and represents an advance over direct-only imaging [2, 9], which requires at least two images. Third, we show that *any* ensemble of structured-light patterns can be made robust to indirect light, regardless of the patterns' frequency content. This involves simply switching from conventional to SLT imaging—without changing the patterns or the algorithm that processes them. As such, our work stands in contrast to prior work on transport-robust structured light, which places the onus on the design of the patterns themselves [6, 12–14]. Fourth, we show that SLT imaging can turn any multi-pattern 3D structured-light method into a one-shot technique for dynamic shape capture. Thus an entire family of previously-inapplicable techniques can be brought to bear on this much-studied problem [5, 15–18] in order to improve depth map resolution and robustness to indirect light. As a proof of concept, we demonstrate in Figure 9 the reconstruction of dense depth and albedo from individual frames of monochrome video, acquired by combining indirect-invariant SLT imaging and conventional six-pattern phase-shifting.

Conceptually, our work has one essential difference from conventional structured light [2, 5]: instead of controlling light only at its source by projecting patterns, we control light at its destination as well, with a DMD mask in front of the camera pixels. This simultaneous projection and masking makes it possible to analyze light transport *geometrically* (by blocking 3D light paths), rather than *photometrically* (by blocking certain transport frequencies and assuming constrained scene reflectance [2]). It also enables optical-domain implementations, which can have a significant speed and signal-to-noise ratio advantage over post-capture processing. The idea was first used in [9] for static scenes and a coaxial projector/camera, where epipolar geometry is degenerate and stereo is impossible. While SLT imaging builds on that work, its premise, theory, applications, and physical implementation are different.

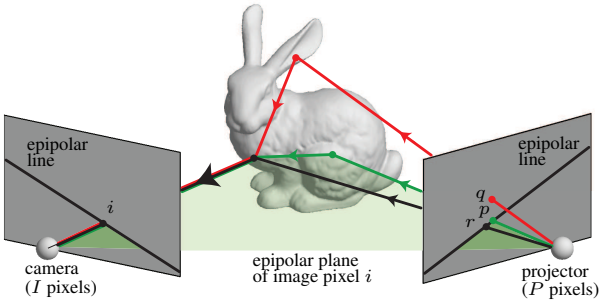

Figure 2: Light transport in a stereo projector-camera system. Light can reach pixel i on the image in one of three general ways: by indirect transport from an arbitrary pixel p on the corresponding epipolar line (green path); by indirect transport from a pixel q that is *not* on that line (red path); or by direct surface reflection, starting from projector pixel r on the epipolar line (black path).

2. The Stereo Transport Matrix

We begin by relating scene geometry to the light transported from a projector to a camera in general position. Consider a scene whose shape potentially varies with time. If the camera and projector respond linearly to light, the scene's instantaneous image satisfies the light transport equation [19]:

$$\mathbf{i} = \mathbf{T}\,\mathbf{p} \tag{1}$$

where $\mathbf{i}$ is the image represented as a column vector of I pixels; $\mathbf{p}$ is the P-pixel projected pattern, also represented as a column vector; and $\mathbf{T}$ is the scene's $I \times P$ instantaneous light transport matrix.

Intuitively, element $\mathbf{T}[i, p]$ of the transport matrix specifies the total radiance transported from projector pixel p to image pixel i over all possible paths. As such, $\mathbf{T}$ models image formation in very general settings: the scene may have non-Lambertian reflectance, it may scatter light volumetrically, exhibit specular inter-reflections, *etc*.

Anatomy of the stereo transport matrix Since a projector and a camera in general position define a stereo pair, their transport matrix is best understood by taking two-view geometry into account. More specifically, we classify the elements of $\mathbf{T}$ into three categories based on the geometry of their transport paths (Figure 2):

- **Epipolar elements,** whose projector and camera pixels are on corresponding epipolar lines. These are the only elements of $\mathbf{T}$ whose transport paths begin and end on rays that can intersect in 3D. By performing stereo calibration [20] and vectorizing patterns and images according to Figure 3, these elements can be made to occupy a known, time-invariant, block-diagonal subset of the transport matrix.

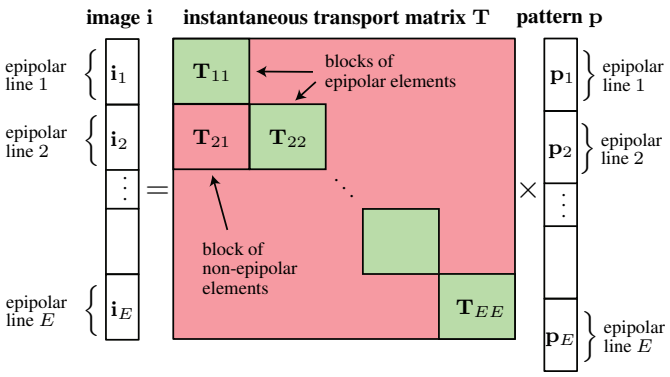

Figure 3: The light transport equation when patterns and images are vectorized so that consecutive pixels on corresponding epipolar lines form subvectors $\mathbf{p}_e$ and $\mathbf{i}_e$, respectively. Under this vectorization scheme, block $\mathbf{T}_{ef}$ of the transport matrix describes transport from epipolar line f on the pattern to epipolar line e on the image. Blocks $\mathbf{T}_{ee}$, shown in green, contain the epipolar elements.

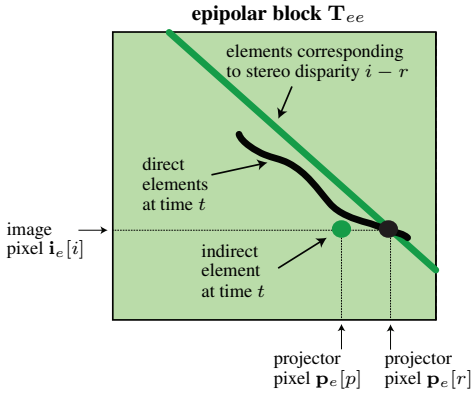

Figure 4: Structure of an epipolar block $\mathbf{T}_{ee}$. Element $\mathbf{T}_{ee}[i, r]$ describes transport from projector pixel $\mathbf{p}_e[r]$ to image pixel $\mathbf{i}_e[i]$. This element is direct if and only the scene point projecting to both pixels is the same, *i.e.*, the point's stereo disparity is $i - r$. The set of direct elements therefore represents the scene's instantaneous disparity map. Conventional stereo algorithms attempt to localize this set while assuming that the transport matrix is zero everywhere else—both inside and outside its epipolar blocks.

- **Non-epipolar elements**, whose projector pixel and camera pixel are not on corresponding epipolar lines. Non-epipolar elements are significant because they vastly outnumber the other elements of $\mathbf{T}$ and *never* account for direct transport. This is because their transport paths begin and end with rays that do not intersect, so light must bounce at least twice to follow them.
- **Direct elements,** whose camera and projector pixels are in stereo correspondence, *i.e.*, they are the perspective projections of a visible surface point. Direct elements are where direct surface reflection actually occurs in the scene; although they always lie within $\mathbf{T}$'s epipolar blocks, their precise location is scene dependent and thus unknown. Indeed, locating the direct elements is equivalent to computing the scene's instantaneous stereo disparity map (Figure 4).

We can therefore express every image of the scene as a sum of three components that arise from distinct "slices" of the transport matrix:

$$\mathbf{i} = \underbrace{\mathbf{T}^{\mathrm{D}}\mathbf{p}}_{\text{direct image}} + \underbrace{\mathbf{T}^{\mathrm{EI}}\mathbf{p}}_{\substack{\text{epipolar}\\ \text{indirect image}}} + \underbrace{\mathbf{T}^{\mathrm{NE}}\mathbf{p}}_{\substack{\text{non-epipolar}\\ \text{indirect image}}} \qquad (2)$$

where the $I \times P$ matrices $\mathbf{T}^{\mathrm{D}}$, $\mathbf{T}^{\mathrm{EI}}$ and $\mathbf{T}^{\mathrm{NE}}$ hold the direct, epipolar indirect, and non-epipolar elements, respectively, and are zero everywhere else.

3. Dominance of Non-Epipolar Transport

Although in theory all three image components in Eq. (2) may contribute to scene appearance, in practice their contributions are not equal. The key observation underlying our work is that the non-epipolar component is very large relative to the epipolar indirect for a broad range of scenes:

$$\mathbf{i} \approx \underbrace{\mathbf{T}^{\mathrm{D}}\mathbf{p}}_{\text{direct image}} + \underbrace{\mathbf{T}^{\mathrm{NE}}\mathbf{p}}_{\text{non-epipolar indirect image}} . \qquad (3)$$

We call this the *non-epipolar dominance assumption*. The transport matrix is much simpler when this assumption holds because we can treat it as having a time-invariant structure with two easily-identifiable parts: the epipolar blocks, which contribute only to the direct image, and the non-epipolar blocks, which contribute only to the indirect.

To motivate this assumption on theoretical grounds, we prove that it holds for two very general scene classes: (1) scenes whose transport function is measurable everywhere and (2) generic scenes containing pure specular reflectors and transmitters. These two cases can be thought of as representing opposite extremes, with the former covering low-frequency transport phenomena such as diffuse interreflection and diffuse isotropic subsurface scattering [21] and the latter covering transport whose frequency content is not band limited. In particular, we prove the following:

Proposition 1. *If $\mathbf{T}$ is the discretized form of a transport function that is measurable and positive over the rectified projector and image planes, then*

$$\lim_{\epsilon \to 0} \frac{\mathbf{T}^{\mathrm{EI}}\mathbf{p}}{\mathbf{T}^{\mathrm{NE}}\mathbf{p}} = \mathbf{0} \qquad (4)$$

where division is entrywise and ϵ is the pixel size for discretization.

Proposition 2. *Two generic n-bounce specular transport paths that originate from corresponding epipolar lines do not intersect for $n > 1$.*

See [8] for proofs. Intuitively, both propositions are consequences of a "dimensionality gap": the set of transport paths contributing to the epipolar indirect image has lower dimension than the set of paths contributing to the non-epipolar image (Figure 2). Thus contributions accumulated in one image are negligible relative to the other in generic settings.

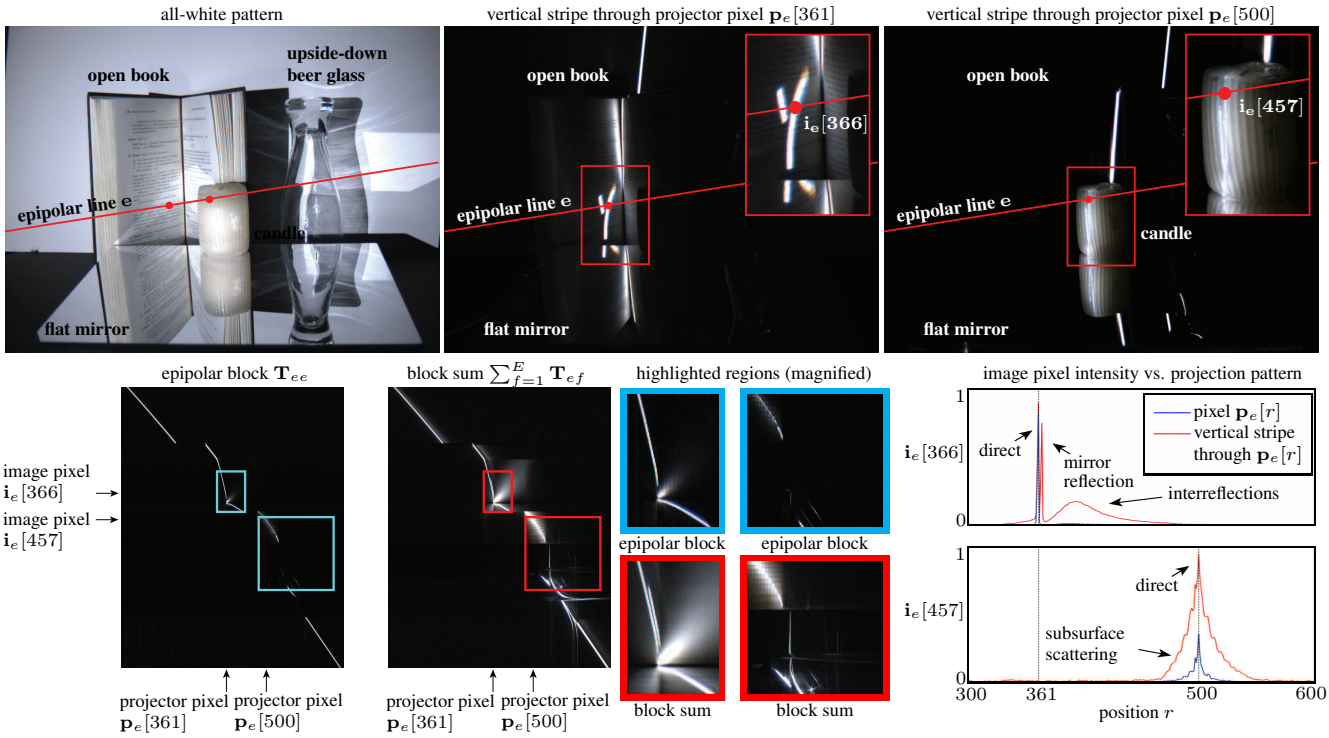

Figure 5: Experimental validation of non-epipolar dominance for a scene containing diffuse, translucent, refractive and mirror-like objects. **Top left:** View under an all-white projection pattern. **Top middle:** View when just one white vertical stripe is projected onto the scene. The many bright regions in this image occur because the stripe illuminates the book's pages in three different ways: (1) directly from the projector, (2) by diffuse inter-reflection from the opposite page, and (3) by specular reflection via the mirror. Their existence makes the scene hard to reconstruct with conventional techniques such as laser-stripe 3D scanning [4]. A magnified view of these regions is shown in the inset. **Top right:** View for another vertical stripe, part of which falls on the candle. The stripe appears very broad and poorly localized there, because of strong sub-surface scattering. **Bottom left:** The epipolar block $\mathbf{T}_{ee}$ for epipolar line e. We show $\mathbf{T}_{ee}$ using the conventions of Figure 4, *i.e.*, its r-th column comes from an image of the scene acquired with only projector pixel $\mathbf{p}_e[r]$ turned on. **Bottom middle:** To assess the image contribution of non-epipolar transport, we acquire the block sum $\sum_{f=1}^{E} \mathbf{T}_{ef}$ and compare it to block $\mathbf{T}_{ee}$—observe that non-epipolar contributions indeed far surpass the epipolar indirect ones. To acquire the block sum, we capture images of the scene while sweeping a vertical stripe on the projector plane (see [8] for a video of the captured image sequence). The r-th column of the block sum is given by the pixels on epipolar line e when the stripe is at $\mathbf{p}_e[r]$. **Bottom right:** Horizontal cross-section of $\mathbf{T}_{ee}$ and $\sum_{f=1}^{E} \mathbf{T}_{ef}$ for two image pixels. Observe that $\mathbf{T}_{ee}$'s cross-section (blue) is sharp and unimodal whereas the block sum's (red) is trimodal for one pixel and very broad for the other.

On the practical side, we have found non-epipolar dominance to be applicable quite broadly; see Figure 5 for a detailed analysis of non-epipolar dominance in a complex scene, Figure 9 for more examples, and [8] for videos confirming the assumption's validity in a variety of settings.

4. Imaging by Structured Light Transport

The rich structure of the stereo transport matrix cannot be exploited by simply projecting a pattern onto the scene. This is because projection gives no control over how light flows through the scene: all elements of $\mathbf{T}$—regardless of position—will participate in image formation. To make full use of $\mathbf{T}$'s structure, *we structure the flow of light itself.*

Our starting point is an imaging procedure first proposed by O'Toole *et al.* [9]. Its main advantage is that the contribution of individual elements of $\mathbf{T}$ can be weighted according to a user-defined "probing matrix" $\mathbf{\Pi}$:

$$\mathbf{i} \;=\; [\,\mathbf{\Pi} \circ \mathbf{T}\,]\,\mathbf{1} \tag{5}$$

where $\circ$ denotes entrywise (*a.k.a.* Hadamard) product and $\mathbf{1}$ is a column vector of all ones. Images captured this way are said to be the result of *probing* the scene's transport matrix with matrix $\mathbf{\Pi}$. Conceptually, they correspond to images of a scene that is illuminated by an all-white pattern and whose transport matrix is $\mathbf{\Pi} \circ \mathbf{T}$.

Two basic questions arise when considering Eq. (5) for image acquisition and shape recovery: (1) what should $\mathbf{\Pi}$ be, and (2) how to design an imaging system that implements the equation? The answers in [9] were restricted to static scenes and projector/camera arrangements that share a single viewpoint, none of which apply here. Below we focus on the first question—designing $\mathbf{\Pi}$—and discuss live imaging of dynamic scenes in Section 5.

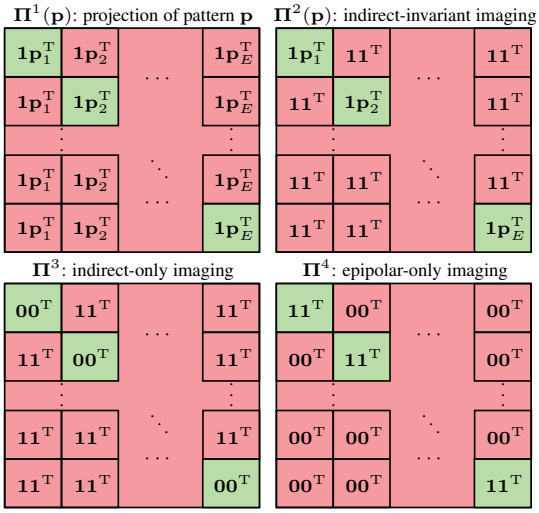

Figure 6: The four basic probing matrices used in this paper. Their block structure mirrors the structure of $\mathbf{T}$ in Figure 3.

Conventional structured-light imaging To gain some insight, let us re-cast as a probing operation the act of projecting a fixed pattern $\mathbf{p}$ and capturing an image $\mathbf{i}$. Applying the vectorization scheme of Figure 3 to the light transport equation and re-arranging terms we get for epipolar line e:

$$\mathbf{i}_e = \sum_{f=1}^{E} \mathbf{T}_{ef}\,\mathbf{p}_f = \Big[\sum_{f=1}^{E} \underbrace{(\mathbf{1}\mathbf{p}_f^{\mathrm{T}})}_{\substack{\text{block of}\\ \text{probing matrix}}} \circ \underbrace{\mathbf{T}_{ef}}_{\text{block of } \mathbf{T}}\Big]\,\mathbf{1} \qquad (6)$$

where E is the number of epipolar lines. Equation (6) implies that projecting $\mathbf{p}$ is equivalent to probing with the matrix $\mathbf{\Pi}^1(\mathbf{p})$ shown in Figure 6. Observe that if we capture images for a whole sequence of projection patterns—as is often the case in structured-light systems—the non-epipolar blocks of the probing matrix will be different for each pattern. Indirect transport will therefore contribute to each captured image differently, and in a way that strongly depends on the particular pattern. This makes structured-light 3D scanning difficult when indirect transport is present because its contributions cannot be easily identified and removed.

Indirect-invariant imaging The contribution of indirect transport becomes much easier to handle if we ensure it is *the same for every pattern*. Since this contribution is dominated by the non-epipolar blocks of the transport matrix, we can achieve (almost) complete invariance to indirect transport by probing with a matrix whose non-epipolar blocks are independent of $\mathbf{p}$. In particular, probing with the matrix $\mathbf{\Pi}^2(\mathbf{p})$ in Figure 6 yields

$$\mathbf{i}_e = \underbrace{\Big[(\mathbf{1}\mathbf{p}_e^{\mathrm{T}}) \circ \mathbf{T}_{ee}\Big]\mathbf{1}}_{\text{direct image (depends on } \mathbf{p})} + \underbrace{\Big[\sum_{f=1, f\neq e}^{E} \mathbf{T}_{ef}\Big]\mathbf{1}}_{\substack{\text{non-epipolar indirect}\\ \text{image (ambient)}}} . \qquad (7)$$

The image in Eq. (7) has two properties: (1) its direct component is identical to the direct component we would get by projecting $\mathbf{p}$ conventionally onto the scene, and (2) its non-epipolar component is independent of $\mathbf{p}$. This independence essentially turns indirect contributions into an "ambient light" term that does not originate from the projection pattern.[1] To see the practical significance of this independence, the second row of Figure 9 compares views of a scene under conventional and one-shot indirect-invariant imaging, for the same projection pattern.

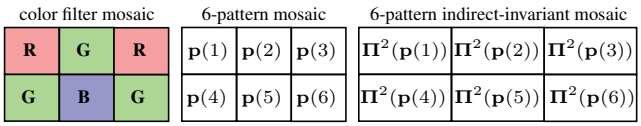

Figure 7: Example layouts for color RGB, monochrome 6-pattern, and monochrome 6-pattern indirect-invariant imaging.

An important corollary of Eq. (7) is that indirect-invariant images can be acquired for *any* sequence of patterns—regardless of frequency content or other properties—using the corresponding sequence of probing matrices.

Indirect-only imaging A notable special case of indirect-invariant imaging is to set $\mathbf{p}$ to zero (matrix $\mathbf{\Pi}^3$ in Figure 6). This yields an image guaranteed to have no contributions from direct transport. Moreover, almost all indirect light will be recorded when non-epipolar dominance holds.

Epipolar-only imaging The exact opposite effect can be achieved with a probing matrix that is zero everywhere except along the epipolar blocks (matrix $\mathbf{\Pi}^4$ in Figure 6). When non-epipolar dominance holds, images captured this way can be treated as (almost) purely direct.

One-shot, multi-pattern, indirect-invariant imaging All four probing matrices in Figure 6 produce views of the scene under a fixed illumination pattern $\mathbf{p}$. With probing, however, it is possible to capture—in just one shot—spatially-multiplexed views of the scene for a whole sequence of structured-light patterns, $\mathbf{p}(1), \ldots, \mathbf{p}(S)$. The probing matrix to achieve this can be thought of as defining a "projection pattern mosaic," much like the RGB filter mosaic does for color (Figure 7). Moreover, we can confer invariance to indirect light by defining the mosaic in terms of *probing matrices* rather than conventional patterns.

Specifically, suppose we partition the I image pixels into S sets and let $\mathbf{b}(1), \ldots, \mathbf{b}(S)$ be binary vectors of size I indicating the pixel membership of each set. The matrix

$$\mathbf{\Pi}^5(\mathbf{p}(1), \ldots, \mathbf{p}(S)) = \sum_{s=1}^{S} \Big[\,\mathbf{b}(s)\,\mathbf{1}^{\mathrm{T}}\,\Big] \circ \mathbf{\Pi}^2(\mathbf{p}(s)) \qquad (8)$$

interleaves the rows of S indirect-invariant probing matrices. Thus, probing with this matrix yields an image containing S sub-images, each of which is a view of the scene under a specific structured-light pattern in the sequence.

[1] Other examples of ambient terms with identical behavior include image contributions from the projector's black level and contributions from light sources other than the projector. Because such terms are often unavoidable yet easy to handle, many structured-light algorithms are designed to either recover them explicitly or be robust to their existence [5]. Non-zero ambient terms do, however, reduce contrast and may affect SNR.

5. Live Structured-Light-Transport Imaging

The feasibility of probing comes from re-writing Eq. (5) as a bilinear matrix-vector product [9]:

$$\mathbf{i} \;=\; \sum_{t=1}^{T} \mathbf{m}(t) \circ [\, \mathbf{T}\, \mathbf{q}(t) \,] \qquad (9)$$

where the transport matrix $\mathbf{T}$ is constant in time and $\mathbf{\Pi} = \sum_{t=1}^{T} \mathbf{m}(t)(\mathbf{q}(t))^{\mathrm{T}}$ is a rank-1 decomposition of the probing matrix. According to Eq. (9), optical probing is possible by (1) opening the camera's shutter, (2) projecting pattern $\mathbf{q}(t)$ onto the scene, (3) using a semi-transparent pixel mask $\mathbf{m}(t)$ to modulate the light arriving at individual camera pixels, (4) changing the pattern and mask synchronously T times, and (5) closing the shutter. This procedure acquires one image; it was implemented in [9] for low-resolution probing matrices using an LCD panel for pixel masking, an SLR camera for image acquisition, and $T \in [100, 1000]$.

Although results were promising, LCDs are not suitable for video-rate (30Hz) probing: they refresh at 30-200Hz, limiting T to an unusable 1-6 masks/projections per frame; and they have low transmittance, requiring long exposure times.

Our approach, on the other hand, is to use a pair of off-the-shelf digital micro-mirror (DMD) devices for projection and masking (Figure 8). These devices are compact, incur no light loss and can operate synchronously at $2.7 - 24$kHz. To implement Eq. (9), we couple them with a conventional video camera operating at 28fps. This allows $96 - 800$ masks/projections within the 36msec exposure of each frame.[2] To our knowledge, such a coupling has not been proposed before.[3]

A major difference between LCDs and DMDs is that DMDs are *binary*. This turns the derivation of masks and projection patterns into a combinatorial optimization problem. Formally, given an *integer*[4] probing matrix $\mathbf{\Pi}$ and an upper bound on T, we seek a length-T rank-1 decomposition into binary vectors such that the decomposition approximates $\mathbf{\Pi}$ as closely as possible. This problem is difficult and we know of no general solution. Indeed, estimating the length of the shortest *exact* decomposition is itself NP-hard [23].

Our approach, below, is to derive randomized decompositions of $\mathbf{\Pi}$ that approximate Eq. (9) in expectation. Although our experience is that this approach works well in practice, it should not be treated as optimal.

Indirect-only imaging Matrix $\mathbf{\Pi}^3$ is a special case where short decompositions are easy. Let $\mathbf{q}(e)$ be a pattern whose pixels are 1 along epipolar line e and 0 everywhere else and let $\mathbf{m}(e)$ be a mask that is 1 everywhere except at epipolar line e. Then it is easy to show that $\mathbf{\Pi}^3 = \sum_{e=1}^{E} \mathbf{m}(e)(\mathbf{q}(e))^{\mathrm{T}}$. This corresponds to a sequence of mask/projection pairs where only one epipolar line is "off" in the mask and only the corresponding epipolar line is "on" in the pattern. Even though this decomposition is exact—and feasible for near-megapixel images—it has poor light efficiency because only one epipolar line is "on" at any time. To improve light efficiency we use random patterns instead, which yield good approximations that are much shorter.

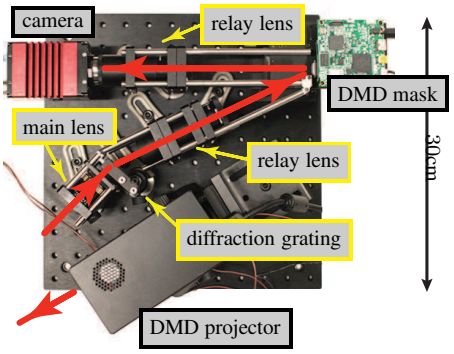

Figure 8: Photo of our prototype. The projector can be detached to change the stereo baseline. The optical path is shown in red. See [8] for a detailed list of components.

Specifically, consider the random pattern

$$\mathfrak{q} = \{\text{each epipolar line is 1 with probability } 0.5\} \;, \qquad (10)$$

let the projection pattern $\mathbf{q}(t)$ be a sample of $\mathfrak{q}$, and let the mask $\mathbf{m}(t)$ be equal to $\overline{\mathbf{q}(t)}$. Taking expectations in Eq. (9), the epipolar line e of the expected image is given by

$$\mathcal{E}[\mathbf{i}_e] \;=\; \mathcal{E}[\overline{\mathfrak{q}_e}] \circ \sum_{\substack{f=1 \\ f \neq e}}^{E} \mathbf{T}_{ef}\, \mathcal{E}[\mathfrak{q}_f] \;=\; 0.25 \sum_{\substack{f=1 \\ f \neq e}}^{E} \mathbf{T}_{ef}\, \mathbf{1} \qquad (11)$$

where $\mathcal{E}[]$ denotes expectation. This is the result of probing with matrix $\mathbf{\Pi}^3$, albeit at one quarter of the "ideal" image intensity.[5] Note that corresponding epipolar lines are never on at the same time in the pattern and mask; thus no epipolar transport path ever contributes to the captured image.

Epipolar-only imaging Matrix $\mathbf{\Pi}^4$ is a special case at the other extreme, where *no* short rank-1 decompositions exist. Since $\mathbf{\Pi}^4 = \mathbf{\Pi}^1(\mathbf{1}) - \mathbf{\Pi}^3$, we compute the result of probing with $\mathbf{\Pi}^4$ by subtracting two adjacent video frames—one captured by projecting an all-white pattern and one captured by indirect-only imaging. Naturally, two-frame motion estimation may be necessary to handle fast-moving scenes (but we do not estimate motion in our experiments).

Indirect-invariant imaging A perhaps counterintuitive result is that even though epipolar-only imaging requires two frames, indirect-invariant imaging requires just one. This is important because probing with matrix $\mathbf{\Pi}^2()$ is all we need

[2] See [9] for an analysis of the SNR advantage confered by performing T mask/projection operations in a single exposure versus capturing an image for each projection pattern, at $1/T$-th the exposure.

[3] The closest design we are aware of comes from confocal microscopy [22]. Its optical path was less challenging to implement, however, because imaging was both coaxial and orthographic.

[4] Since any grayscale structured-light pattern $\mathbf{p}$ must be quantized before projection, probing matrices are always integer, including $\mathbf{\Pi}^2(\mathbf{p})$.

[5] Intuitively, since half the epipolar lines are "off" in the pattern and the mask, only 1/4th of the total light is transported from projector to camera.

for reconstruction with structured light. Let $\mathbf{p}$ be an arbitrary structured-light pattern scaled to $[0, 1]$. Define mask $\mathbf{m}(t)$ to be a sample of $\mathfrak{q}$ from Eq. (10) and the pattern to be

$$\mathbf{q}(t) \;=\; \mathbf{m}(t) \circ \mathbf{r}(t) \;+\; \overline{\mathbf{m}(t)} \circ \overline{\mathbf{r}(t)} \tag{12}$$

where $\mathbf{r}(t)$ is a sample of yet another random pattern:

$$\mathfrak{r} = \{\text{pixel } p \text{ on epipolar line } e \text{ is 1} \\ \text{with probability } \mathbf{p}_e[p]\} \;. \tag{13}$$

A pictorial illustration of Eq. (12) can be found in [8]. From calculations similar to Eq. (11), the expected image is

$$\begin{aligned} \mathcal{E}[\mathbf{i}_e] \;&=\; 0.5\mathbf{T}_{ee}\mathbf{p}_e + 0.25 \sum_{f=1, f\neq e}^{E} [\mathbf{T}_{ef}\mathbf{p}_f + \mathbf{T}_{ef}(\mathbf{1} - \mathbf{p}_f)] \\ &=\; \underbrace{0.5\mathbf{T}_{ee}\mathbf{p}_e}_{\substack{\text{direct image} \\ \text{(depends on } \mathbf{p}\text{)}}} \;+\; \underbrace{0.25 \sum_{f=1, f\neq e}^{E} \mathbf{T}_{ef}\mathbf{1}}_{\text{indirect image (ambient)}} \;, \end{aligned} \tag{14}$$

which is equivalent to the result of probing with $\mathbf{\Pi}^2()$.

One-shot, multi-pattern, indirect-invariant imaging Here we use the mask for indirect-invariant imaging and temporally multiplex S random projection patterns—each defined by Eq. (12) and corresponding to a different structured-light pattern— across our "budget" of T total projections per video frame. After the video is recorded, we "demosaic" each frame $\mathbf{i}$ independently to infer S full-resolution images, one for each structured-light pattern. Following work on compressed sensing [24, 25] we do this by solving for S images that reproduce frame $\mathbf{i}$ and are sparse under a chosen basis $\mathbf{W}$:

$$\text{minimize} \quad \left\| \mathbf{W}^{\mathbf{T}} \begin{bmatrix} \mathbf{i}(1) & \dots & \mathbf{i}(S) \end{bmatrix} \right\|_n \tag{15}$$

$$\text{subject to} \quad \left\| \sum_{s=1}^{S} \mathbf{b}(s) \circ \mathbf{i}(s) - \mathbf{i} \right\|_2 \leq \epsilon \tag{16}$$

where $\|.\|_n$ is a sparsity-inducing norm[6] and $\mathbf{b}(s)$ is the binary vector holding pixel memberships for pattern s.

6. Experimental Results

Indirect-only and epipolar-only imaging Our DMDs operated at 2.7kHz with $T = 96$ or 48 patterns/masks per frame. For calibration, we computed the epipolar geometry between the two DMDs by first relating them to the image plane. Overall resolution was equal to the resolution of our DMDs, *i.e.*, 608×684. See Figures 1 and 9 (row 1) for examples of indirect- and epipolar-only images, respectively.

Indirect-invariant imaging We used high-end DMDs and a monochrome camera for the reconstruction experiments in Figure 9 (rows 2 and 3), with $T = 800$ patterns/masks per frame. The effective DMD resolution was approximately 484×364. The scenes occupied a $40^3 cm^3$ volume about $70cm$ away from the camera. To show the effectiveness of SLT imaging, we chose the most basic pattern and technique—phase-shifting with 9 sinusoids total, at frequencies 1, 8 and 64.

Dense depth and albedo from one shot We used $S = 6$ sinusoids at frequencies 4 and 32 for the experiment in Figure 9 (row 4), and a random, rather than regular, assignment of pixels to sinusoids. We recorded multi-pattern, indirect-invariant video at 28fps and reconstructed each frame independently by (1) solving for the 6 demosaiced patterns using SPGL1 [26] for optimization and the JPEG2000 wavelet basis, and (2) using them to get per-pixel depth and albedo.

7. Concluding Remarks

We believe that optical-domain processing—and SLT imaging in particular—offers a powerful new way to analyze the appearance of complex scenes, and to boost the abilities of existing reconstruction algorithms. Although our focus was mainly on monochromatic light and conventional cameras, SLT imaging depends on neither; integrating this framework with other imaging dimensions (polarization, wavelength, time, *etc*.) is a promising direction. Last but not least, although our prototypes rely on DMD masks and several optical components, these would be rendered unnecessary if per-pixel processing was implemented directly on the sensor [27, 28]. We are looking forward to the wide availability of such technologies.

Acknowledgements We are grateful for the support of the Natural Sciences and Engineering Research Council of Canada under the PGS-D, RGPIN, RTI, SGP and GRAND NCE programs.

References

[1] P. Debevec, *et al*. Acquiring the reflectance field of a human face. *Proc. SIGGRAPH'00*.
[2] S. K. Nayar, *et al*. Fast separation of direct and global components of a scene using high frequency illumination. *Proc. SIGGRAPH'06*.
[3] G. Godin, M. Rioux, J. Beraldin, and M. Levoy. An assessment of laser range measurement on marble surfaces. *Proc. 5th Conf. on Optical 3D Measurement Techniques*, 2001.
[4] B. Curless and M. Levoy. Better optical triangulation through spacetime analysis. *Proc. ICCV'95*.
[5] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado. A state of the art in structured light patterns for surface profilometry. *Pattern Recogn*, 43(8):2666–2680, 2010.
[6] M. Gupta, S. Nayar. Micro Phase Shifting. *Proc. CVPR'12*.
[7] S. K. Nayar, K. Ikeuchi, and T. Kanade. Shape from Interreflections. *Int. J. Computer Vision*, 6(3):173–195, 1991.
[8] M. O'Toole, J. Mather, and K. N. Kutulakos. Supplementary materials. *http://www.dgp.toronto.edu/~motoole/slt*.
[9] M. O'Toole, R. Raskar, K. N. Kutulakos. Primal-dual coding to probe light transport. *Proc. SIGGRAPH'12*.
[10] A. Velten, *et al*. Femto-photography: capturing and visualizing the propagation of light. *Proc. SIGGRAPH'13*.
[11] S. Achar, S. T. Nuske, S. G. Narasimhan. Compensating for Motion During Direct-Global Separation. *Proc. ICCV'13*.
[12] M. Gupta, *et al*. Structured light 3D scanning in the presence of global illumination. *Proc. CVPR'11*.
[13] V. Couture, N. Martin, and S. Roy. Unstructured light scanning to overcome interreflections. *Proc. ICCV'11*.

[6] We use the $(1, 2)$-norm because it promotes group sparsity and thus concentrates non-zero terms to the same pixels across views.

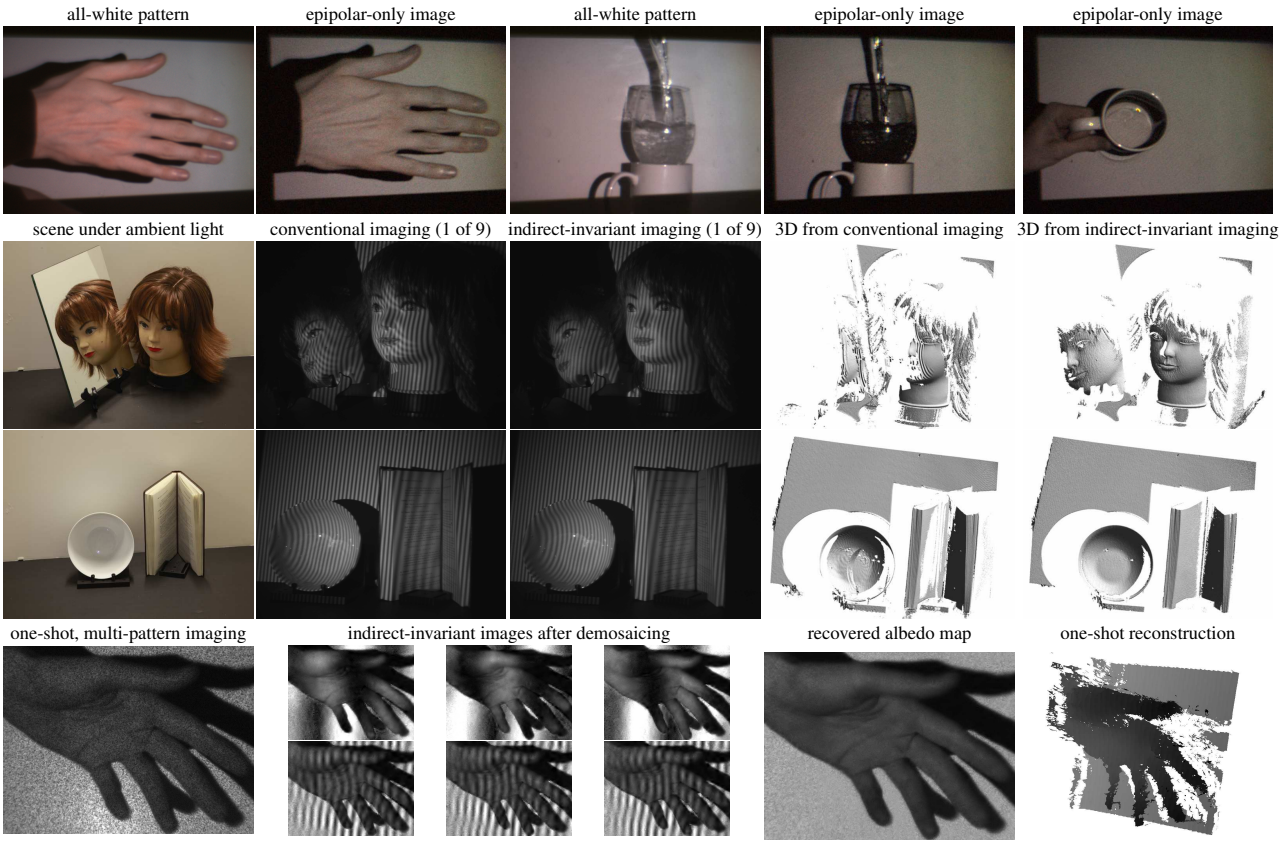

Figure 9: **Row 1:** Frames from conventional and epipolar-only video for the scenes in Figure 1. Compared to Figure 1, the water's opaque appearance in the epipolar-only frame and the absence of caustics in the mug confirm that significant indirect light was not recorded, *i.e.*, non-epipolar dominance holds. Refer to [8] for videos of several more scenes. **Row 2:** We imaged the scene on the left in two ways: (1) projecting 9 phase-shifted patterns directly onto it and (2) capturing indirect-invariant images for the same patterns. Exposure time was held fixed, giving an SNR advantage to conventional projection which does not mask pixels. We then applied the same algorithm to the two sets of images, with the results shown on the right. The algorithm fails catastrophically for the conventionally-acquired images whereas with SLT imaging it is able to reconstruct even the hidden side of the face, from the mirror's indirect view. Closer inspection of the input images (please zoom in) reveals the reason for the difference: the conventional image contains "double fringes" from secondary reflections whereas the indirect-invariant one does not. **Row 3:** Another example, for a scene with strong diffuse and specular inter-reflections. **Row 4:** Reconstructing dense depth and albedo from one video frame of a moving hand. From this frame, our demosaicing algorithm recovers 6 full-resolution indirect-invariant images of the hand, for 6 sinusoidal patterns. These images yield the albedo and depth maps on the right.

[14] T. Chen, H.-P. Seidel, and H. P. A. Lensch. Modulated phase-shifting for 3D scanning. *Proc. CVPR'08*.

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

[16] H. Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi. Dynamic scene shape reconstruction using a single structured light pattern. *Proc. CVPR'08*.

[17] C. Hernandez, *et al*. Non-rigid Photometric Stereo with Colored Lights. *Proc. ICCV'07*.

[18] G. Fyffe, X. Yu, and P. Debevec. Single-shot photometric stereo by spectral multiplexing. *Proc. ICCP'11*.

[19] R. Ng, R. Ramamoorthi, and P. Hanrahan. All-frequency shadows using non-linear wavelet lighting approximation. *Proc. SIGGRAPH'03*.

[20] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Dec. 2000.

[21] H. Jensen, *et al*. A practical model for subsurface light transport. *Proc. SIGGRAPH'01*.

[22] R. Heintzmann, *et al*. A dual path programmable array microscope (PAM). *J. Microscopy*, 204:119–135, 2001.

[23] J. Zhong. Binary ranks and binary factorizations of nonnegative integer matrices. *Electron. J. Linear Algebra*, 23:540–552, 2012.

[24] D. Reddy, A. Veeraraghavan, and R. Chellappa. P2C2: Programmable pixel compressive camera for high speed imaging. *Proc. CVPR'11*.

[25] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar. Video from a single coded exposure photograph using a learned over-complete dictionary. *Proc. ICCV'11*.

[26] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction. *http://www.cs.ubc.ca/~mpf/spgl1*.

[27] M. W. Kelly and M. H. Blackwell. Digital-pixel FPAs enhance infrared imaging capabilities. *Laser Focus World*, 49(1):90, 2013.

[28] G. Wan, *et al*. CMOS Image Sensors With Multi-Bucket Pixels for Computational Photography. *IEEE J. Solid State Circuits*, 47(4):1031–1042, 2012.

3D Shape and Indirect Appearance By Structured Light Transport: Supplemental Document

Matthew O'Toole John Mather Kiriakos N. Kutulakos
Department of Computer Science
University of Toronto
{motoole,jmather,kyros}@cs.toronto.edu

A. Proofs of Propositions 1 and 2

A.1. Proof of Proposition 1

Proposition 1. *If* $\mathbf{T}$ *is the discretized form of a transport function that is measurable and positive over the rectified projector and image planes, then*

$$\lim_{\epsilon \to 0} \frac{\mathbf{T}^{\mathrm{EI}}\,\mathbf{p}}{\mathbf{T}^{\mathrm{NE}}\,\mathbf{p}} \;=\; \mathbf{0} \tag{17}$$

where division is entry-wise and ϵ *is the pixel size for discretization.*

Proof sketch. We begin by identifying the rectified projector and image planes with the continuous domain $\mathcal{D} = [-1,1] \times [-1,1] \subset \Re^2$. Let $p = (p_x, p_y)$ be a point on the projector plane and let $\mathbf{I}_\epsilon(p)$ be an indicator function over $\mathcal{D}$ that specifies the spatial extent of the discrete epipolar line through the origin:

$$\mathbf{I}_\epsilon(p) = \begin{cases} 1 & \text{if } |p_x| \leq \frac{\epsilon}{2} \text{ and } |p_y| \leq 1 \\ 0 & \text{otherwise .} \end{cases} \tag{18}$$

In the continuous setting, light transport from the projector plane to the image plane is described by the *light transport equation* [2]. Given an image point $i \in \mathcal{D}$ on the epipolar line through the origin, this equation describes the total radiance transported to i from points on the projector plane:

$$\mathcal{I}(i) \;=\; \underbrace{\mathcal{T}(\hat{p},\, i)\, \mathcal{P}(\hat{p})}_{\text{direct}} \;+\; \underbrace{\int_{\mathcal{D}-\{\hat{p}\}} \mathcal{T}(p,\, i)\, \mathcal{P}(p)\, dp}_{\text{indirect}} \tag{19}$$

where $\hat{p}$ is the projector point in stereo correspondence with image point i; $\mathcal{P}(p)$ is the radiance along the ray through projector point p; and $\mathcal{T}(p,i)$ is the *transport function* describing the proportion of radiance from p that gets transported to i.

Without loss of generality, we prove the continuous form of the ratio in Eq. (1) for an image point i; this point is taken to be inside a discrete image pixel of dimension $\epsilon \times \epsilon$ on the epipolar line through the origin.

More specifically, we consider the epipolar indirect, total indirect, and non-epipolar indirect contributions at i:

$$\mathcal{I}^{\mathrm{EI}}(i) \;=\; \int_{\mathcal{D}-\{\hat{p}\}} \mathbf{I}_\epsilon(p)\, \mathcal{T}(p,i)\, \mathcal{P}(p)\, dp \tag{20}$$

$$\mathcal{I}^{\mathrm{I}}(i) \;=\; \int_{\mathcal{D}-\{\hat{p}\}} \mathcal{T}(p,i)\, \mathcal{P}(p)\, dp \tag{21}$$

$$\mathcal{I}^{\mathrm{NE}}(i) \;=\; \mathcal{I}^{\mathrm{I}}(i) \;-\; \mathcal{I}^{\mathrm{EI}}(i)\;. \tag{22}$$

We now show that for any $\delta > 0$, there is an $\epsilon > 0$ such that

$$\left| \frac{\mathcal{I}^{\mathrm{EI}}(i)}{\mathcal{I}^{\mathrm{NE}}(i)} \right| \;<\; \delta\;. \tag{23}$$

Since $\mathcal{T}()$ is measurable, we can apply the Cauchy-Schwarz inequality [4] to Eq. (4) to get an upper bound on the epipolar indirect contributions:

$$\begin{aligned} \mathcal{I}^{\mathrm{EI}}(i) \;&\leq\; \left\{ \int_{\mathcal{D}-\{\hat{p}\}} \mathbf{I}_\epsilon(p) \mathrm{d}p \right\}^{\frac{1}{2}} \left\{ \int_{\mathcal{D}-\{\hat{p}\}} \left[\mathcal{T}(p,\, i)\, \mathcal{P}(p) \right]^2 \mathrm{d}p \right\}^{\frac{1}{2}} \\ &=\; (2\epsilon)^{\frac{1}{2}} \left\{ \int_{\mathcal{D}-\{\hat{p}\}} \left[\mathcal{T}(p,\, i)\, \mathcal{P}(p) \right]^2 \mathrm{d}p \right\}^{\frac{1}{2}}\;. \end{aligned} \tag{24}$$

By combining Eqs. (5), (6) and (8) we also get a lower bound on the non-epipolar contributions:

$$\begin{aligned} \mathcal{I}^{\mathrm{NE}}(i) \;\geq\; &\int_{\mathcal{D}-\{\hat{p}\}} \mathcal{T}(p,i)\, \mathcal{P}(p)\, dp \;- \\ &(2\epsilon)^{\frac{1}{2}} \left\{ \int_{\mathcal{D}-\{\hat{p}\}} \left[\mathcal{T}(p,\, i)\, \mathcal{P}(p) \right]^2 \mathrm{d}p \right\}^{\frac{1}{2}}\;. \end{aligned} \tag{25}$$

Equation (7) now follows by choosing ϵ to be

$$\epsilon \;=\; \frac{1}{2} \left(\frac{\delta}{2 \,+\, \delta} \right)^2 \frac{\left\{ \int_{\mathcal{D}-\{\hat{p}\}} \mathcal{T}(p,i)\, \mathcal{P}(p)\, dp \right\}^2}{\int_{\mathcal{D}-\{\hat{p}\}} \left[\mathcal{T}(p,\, i)\, \mathcal{P}(p) \right]^2 \mathrm{d}p}\;. \tag{26}$$

Specifically, substituting Eq. (10) into Eqs. (8) and (9) we get

$$\left| \frac{\mathcal{I}^{\mathrm{EI}}(i)}{\mathcal{I}^{\mathrm{NE}}(i)} \right| \;\leq\; \frac{\frac{\delta}{2+\delta}}{1 - \frac{\delta}{2+\delta}} \;=\; \frac{\delta}{2} \;<\; \delta\;. \tag{27}$$

□

A.2. Proof of Proposition 2

We prove Proposition 2 for scenes consisting of a finite collection of objects, each of which is an open set in $\Re^3$ bounded by a smooth generic surface [1, 3].

Proposition 2. *Two generic n-bounce specular transport paths that originate from corresponding epipolar lines do not intersect for $n > 1$.*

Proof. For simplicity, we reverse the direction of light travel through image pixels, treating the camera as a second projector that also sends light onto the scene.

Let $\mathcal{L}, \mathcal{L}'$ be a pair of corresponding epipolar lines on the (continuous) projector and image planes, respectively, and let $p \in \mathcal{L}$ and $i \in \mathcal{L}'$ be points on them.

Suppose that the light originating at p and i undergoes $n \geq 1$ consecutive specular bounces upon entering the scene. Furthermore, suppose that the associated transport paths are generic, *i.e.*, they remain stable under infinitesimal perturbations of the scene's surfaces and of the points p and i. To prove the proposition, we show that the following cannot hold simultaneously:

1. the transport paths through p and i intersect at their $(n+1)$-th bounce, *i.e.*, their $(n+1)$-th bounce occurs at the same surface point in the scene; and

2. this intersection is generic, *i.e.*, it occurs for all points p, i in an open interval $\mathcal{Q} \subset \mathcal{L}$ and $\mathcal{Q}' \subset \mathcal{L}'$, respectively.

In particular, let $l_n(p)$ be the 3D ray that light follows after n specular bounces from projector point p. Similarly, let $l'_n(i)$ be the corresponding 3D ray for image point i. Since the transport paths through p and i are generic, the mappings $p \mapsto l_n(p)$ and $i \mapsto l'_n(i)$ are smooth functions for some open neighborhood $\mathcal{Q} \subset \mathcal{L}$ and $\mathcal{Q}' \subset \mathcal{L}'$ of p and i, respectively. These mappings define a pair of ruled surfaces in $\Re^3$: intuitively, as point p ranges over $\mathcal{Q}$, the 3D ray $l_n(p)$ twists and translates in space, tracing a ruled surface.

Now, for the transport paths through p and i to have their $(n+1)$-th bounce in common, three surfaces must meet at a point: ruled surface $l_n(\mathcal{Q})$, ruled surface $l'_n(\mathcal{Q}')$, and a surface in the scene. This, however, is *not* a generic condition because surfaces $l_n(\mathcal{Q})$ and $l'_n(\mathcal{Q}')$ transversally intersect along a curve and this curve will transversally intersect the scene's surfaces at isolated points [1]. □

B. Expanded derivations of selected equations

B.1. Derivation of Eq. (11)

Combining Eqs. (6) and (9) we have

$$\mathbf{i}_e \;=\; \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{E} \overline{\mathbf{q}_e(t)} \circ [\, \mathbf{T}_{ef}\; \mathbf{q}_f(t)\,] \tag{28}$$

where the $1/T$ factor captures the fact that each term in the sum is allocated $1/T$ of the total exposure time. We now split the sum into its epipolar and non-epipolar terms

$$\mathbf{i}_e \;=\; \frac{1}{T}\sum_{t=1}^{T}\Bigg\{ \overline{\mathbf{q}_e(t)} \circ [\, \mathbf{T}_{ee}\; \mathbf{q}_e(t)\,] \;+\; \sum_{\substack{f=1\\ f\neq e}}^{E} \overline{\mathbf{q}_e(t)} \circ [\, \mathbf{T}_{ef}\; \mathbf{q}_f(t)\,] \Bigg\} \tag{29}$$

and observe that the first term is always a vector of zeros. Therefore,

$$\mathbf{i}_e \;=\; \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{f=1\\ f\neq e}}^{E} \overline{\mathbf{q}_e(t)} \circ [\, \mathbf{T}_{ef}\; \mathbf{q}_f(t)\,] \;. \tag{30}$$

Letting $T \to \infty$ and applying the Central Limit Theorem to Eq. (14) we get the expected image $\mathcal{E}[\mathbf{i}_e]$ for epipolar line e:

$$\mathcal{E}[\mathbf{i}_e] \;=\; \mathcal{E}\Bigg[\sum_{\substack{f=1\\ f\neq e}}^{E} \overline{\mathbf{q}_e} \circ [\, \mathbf{T}_{ef}\; \mathbf{q}_f\,] \Bigg] \tag{31}$$

$$\;=\; \mathcal{E}[\overline{\mathbf{q}_e}] \;\circ\; \mathcal{E}\Bigg[\sum_{\substack{f=1\\ f\neq e}}^{E} \mathbf{T}_{ef}\; \mathbf{q}_f \Bigg] \tag{32}$$

$$\;=\; \mathcal{E}[\overline{\mathbf{q}_e}] \;\circ\; \sum_{\substack{f=1\\ f\neq e}}^{E} \mathbf{T}_{ef}\; \mathcal{E}[\mathbf{q}_f] \tag{33}$$

$$\;=\; 0.25 \sum_{\substack{f=1\\ f\neq e}}^{E} \mathbf{T}_{ef}\; \mathbf{1} \;, \tag{34}$$

where Eq. (16) follows from the fact that epipolar lines e and f are distinct and thus their corresponding random vectors $\mathbf{q}_e$ and $\mathbf{q}_f$ are independent.

B.2. Derivation of Eq. (14)

Combining Eqs. (6) and (9) for the indirect-invariant mask and pattern we have:

$$\mathbf{i}_e = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{E} \mathbf{m}_e(t) \circ \Big\{ \mathbf{T}_{ef}\; [\, \mathbf{m}_f(t) \circ \mathbf{r}_f(t) + \overline{\mathbf{m}_f(t)} \circ \overline{\mathbf{r}_f(t)}\,] \Big\} \tag{35}$$

We split the sum in Eq. (19) into its epipolar and non-epipolar terms,

$$\mathbf{i}_e \;=\; \frac{1}{T}\sum_{t=1}^{T}\Big\{\mathbf{m}_e(t)\circ\mathbf{T}_{ee}\,[\,\mathbf{m}_e(t)\circ\mathbf{r}_e(t)\,]\;+\;\mathbf{m}_e(t)\circ\mathbf{T}_{ee}\,[\,\overline{\mathbf{m}_e(t)}\circ\overline{\mathbf{r}_e(t)}\,]\;+\;\sum_{\substack{f=1\\f\neq e}}^{E}\mathbf{m}_e(t)\circ\mathbf{T}_{ef}\,[\,\mathbf{m}_f(t)\circ\mathbf{r}_f(t)\,]\;+\;\sum_{\substack{f=1\\f\neq e}}^{E}\mathbf{m}_e(t)\circ\mathbf{T}_{ef}\,[\,\overline{\mathbf{m}_f(t)}\circ\overline{\mathbf{r}_f(t)}\,]\Big\} \quad (36)$$

and note that the second term of Eq. (20) is always a vector of zeros. Letting $T\rightarrow\infty$ and applying the Central Limit Theorem to Eq. (20) we get the expected image for epipolar line e:

$$\mathcal{E}[\mathbf{i}_e] \;=\; \mathcal{E}\Big[\,\mathfrak{q}_e\circ[\,\mathbf{T}_{ee}\,(\mathfrak{q}_e\circ\mathfrak{r}_e)\,]\;+\;\sum_{\substack{f=1\\f\neq e}}^{E}\mathfrak{q}_e\circ[\,\mathbf{T}_{ef}\,(\mathfrak{q}_f\circ\mathfrak{r}_f+\overline{\mathfrak{q}_f}\circ\overline{\mathfrak{r}_f})\,]\,\Big]\;. \quad (37)$$

Now, $\mathfrak{q}_e$ is a random binary vector whose probability of being either $\mathbf{1}$ or $\mathbf{0}$ is 0.5. Using this fact as well as $\mathfrak{q}_e$'s independence from all other random vectors, the expectation in Eq. (21) becomes

$$\mathcal{E}[\mathbf{i}_e] \;=\; 0.5\,\mathbf{T}_{ee}\,\mathcal{E}[\mathfrak{r}_e]\;+\;0.5\sum_{\substack{f=1\\f\neq e}}^{E}\mathbf{T}_{ef}\,\mathcal{E}[\mathfrak{q}_f\circ\mathfrak{r}_f+\overline{\mathfrak{q}_f}\circ\overline{\mathfrak{r}_f}]\;. \quad (38)$$

Finally, using the definition of binary random vector $\mathfrak{r}_f$ in Eq. (13) the expectation becomes

$$\mathcal{E}[\mathbf{i}_e] \;=\; 0.5\,\mathbf{T}_{ee}\,\mathbf{p}_e\;+\;0.5\sum_{\substack{f=1\\f\neq e}}^{E}\mathbf{T}_{ef}\big\{\,\mathrm{Prob}[\,\mathfrak{q}_f=\mathbf{1}\,]\,\mathbf{p}_e+\mathrm{Prob}[\,\mathfrak{q}_f=\mathbf{0}\,]\,(\mathbf{1}-\mathbf{p}_e)\,\big\} \quad (39)$$

which is equal to

$$\mathcal{E}[\mathbf{i}_e] \;=\; 0.5\,\mathbf{T}_{ee}\,\mathbf{p}_e\;+\;0.25\sum_{\substack{f=1\\f\neq e}}^{E}\big\{\mathbf{T}_{ef}(\,\mathbf{p}_e\;+\;(\mathbf{1}-\mathbf{p}_e)\,\big\} \quad (40)$$

$$\;=\; 0.5\,\mathbf{T}_{ee}\,\mathbf{p}_e\;+\;0.25\sum_{\substack{f=1\\f\neq e}}^{E}\mathbf{T}_{ef}\mathbf{1}\;. \quad (41)$$

C. Experimental Prototypes

To encourage reproducibility, we include the complete parts list for our two experimental systems:

- a low-speed, low-cost system for video-rate indirect-only and epipolar-only imaging (Figure 8 of the paper) whose components are listed in Table 1; and
- a high-speed system for indirect-invariant shape acquisition and one-shot multi-pattern imaging (Figure 1), whose components are listed in Table 2.

Indirect-only and epipolar-only imaging We used a color AVT GT1920C camera for acquisition, a Texas Instruments LightCrafter for pixel masking and a 100 lumen Keynote Photonics LightCrafter kit for projection. The DMDs were synchronized at 2.7kHz, permitting $T = 96$ patterns and masks per video frame. The camera and DMD resolutions were quite different—1936x1456 versus 608x684—with each DMD pixel mapping to a 2×2 block of camera pixels. System calibration consists of computing the epipolar geometry between the two DMDs. We did this by first computing correspondences between the camera and each DMD separately. Patterns are uploaded to both DMDs once, at the beginning of an imaging session.

Indirect-invariant imaging For these experiments we used a monochrome AVT GT1920 camera and a pair of high-end DMDs from Texas Instruments (DLi 4130) that use a 2000 lumen light source. These operate at 22.2kHz, permitting $T = 800$ patterns per video frame. Although the DMD resolution was fairly high at 1024×768, its effective resolution was much lower, 484×364, because of the different physical dimensions and orientation of the camera sensor and DMD.

One-shot multi-pattern imaging Effective DMD resolution was even lower, 256×256, because of the scene's limited extent within the camera's field of view.

D. Generation of Masks & Projection Patterns

The mathematical definition of the patterns and masks we use in our SLT prototypes is discussed in Section 5 of the paper. Here, we show in Figure 2 examples of actual mask/pattern pairs uploaded on our DMDs and illustrate their construction according to Eqs. (10), (12), (13) and (16) in the paper.

Indirect-only imaging We use random mask/pattern pairs like those shown in Row 1 of Figure 2. To reduce the sensation of flicker by users who are physically present during video acquisition, we generate a random sequence of $T/2$ mask/pattern pairs ($T = 800$ or 96 depending on the prototype) and then generate a second mask/pattern sequence

Item #	Part Description	Quantity	Model Name	Company
1	color camera	1	GT1920C	Allied Vision Technologies
2	DMD projector	1	LC3000-Pro Pico Projector	Keynote Photonics
3	power supply	1	LC3000-Pro Power Supply	Keynote Photonics
4	DMD projector (mask)	1	DLP LightCrafter	Texas Instruments
5	power supply	1	T1228-Z12P-ND	Digi-Key Corporation
6	connector housing	2	WM1722-ND	Digi-Key Corporation
7	crimp	4	WM1142CT-ND	Digi-Key Corporation
8	Hirose contact plug	1	HR1623-ND	Digi-Key Corporation
9	12mm f/1.4 objective lens	1	Cinegon 1.4/12-0906	Schneider Optics
10	visible achromatic doublet pairs	2	MAP10100100-A	Thorlabs
11	300 grooves/mm transmission grating	1	GT25-03	Thorlabs
12	ring-activated threaded iris diaphragm	2	SM1D12D	Thorlabs
13	C-mount to SM1 adapter	1	SM1A9	Thorlabs
14	SM1 to C-mount adapter	1	SM1A10	Thorlabs
15	SM1 Coupler	1	SM1T10	Thorlabs
16	SM1 Lens Tube, 2 inch Thread Depth	1	SM1L20	Thorlabs
17	SM1 Lens Tube, 3 inch Thread Depth	1	SM1L30	Thorlabs
18	SM1-threaded cage plate	1	CP4S	Thorlabs
19	cage plate with 1.2 inch double bore	5	CP12	Thorlabs
20	cage plate with 35 mm aperture	4	CP03/M	Thorlabs
21	cylindrical lens mount	1	CH1A	Thorlabs
22	rod swivel coupler (set of four)	1	C2A	Thorlabs
23	rod end swivel connector (set of four)	1	C3A	Thorlabs
24	cage assembly rod, 2 inch long	2	ER2	Thorlabs
25	cage assembly rod, 3 inch long	4	ER3	Thorlabs
26	cage assembly rod, 4 inch long	4	ER4	Thorlabs
27	aluminum breadboard	1	MB3030/M	Thorlabs
28	12.7 mm x 40 mm optical post	1	TR40/M	Thorlabs
29	12.7 mm x 100 mm optical post	6	TR100/M	Thorlabs
30	post holder, 40 mm	1	PH40/M	Thorlabs
31	post holder, 100 mm	6	PH100/M	Thorlabs
32	studded pedestal base adapter	7	BE1/M	Thorlabs
33	small clamping fork	7	CF125	Thorlabs
34	30 mm single axis translation stage	1	#66-397	Edmund Optics
35	bottom adapter plate	1	#66-620	Edmund Optics
36	top adapter plate	1	#66-493	Edmund Optics
37	metric base plate	1	#54-975	Edmund Optics
38	thread-to-thread adapter	1	#56-323	Edmund Optics

Table 1: List of parts for the system shown in Figure 8 of the paper. The camera outputs live video at a rate of 28 frames per second. Each video frame requires 96 binary masks/projection patterns, *i.e.*, each mask/pattern is active for $375\mu sec$ of the frame's $36000\mu sec$ total exposure time.

Item #	Part Description	Quantity	Model Name	Company
1	monochrome camera	1	GT1920	Allied Vision Technologies
2	high-speed DMD	2	DLi4130VIS-7XGA	Digital Light Innovations
3	high-power LED	1	High Power S2+ w/ LED	Digital Light Innovations
4	connector housing	2	WM1728-ND	Digi-Key Corporation
5	fixed filter holder 40 mm Sq.	1	#54-997	Edmund Optics
6	45 degree mounting adapter	1	#59-001	Edmund Optics

Table 2: The high-speed system in Figure 1 uses identical optics to the low-speed one; the only differences between the two systems are (1) a faster DMD projector and mask, (2) their mounts, and (3) a monochrome camera that is identical to the low-speed system's color camera, but without the RGB filter mosaic. We only list the differing components in this table; the remaining parts are items 7-38 in Table 1. The system's camera outputs live video at a rate of 28 frames per second. Each video frame consists of 800 binary mask/projection patterns, *i.e.*, each mask/pattern is active for $45\mu sec$ of the frame's $36000\mu sec$ total exposure time.

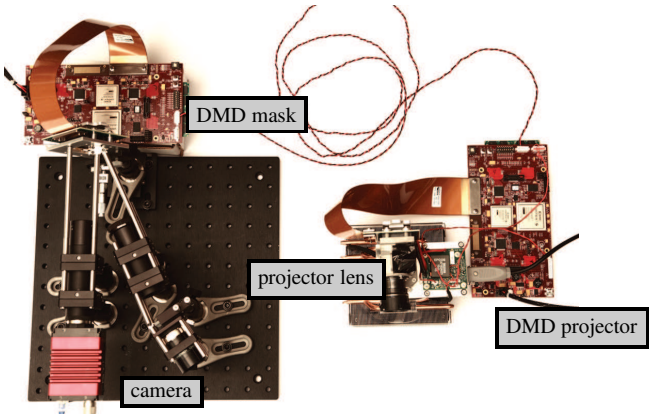

Figure 10: Our high-speed system. The key differences between this system and that shown in Figure 8 of the paper are a monochrome camera, the DMD mask, and the DMD projector.

whose projection patterns are the binary complement of the first $T/2$ projection patterns. This ensures a stable perception because the image integrated by the eye (or by a maskless camera) over the period of one video frame corresponds to a view of the scene under an all-white projection pattern.[1]

For indirect-only imaging, it is also important to ensure that no direct light "leaks" accidentally through the DMD mask. Such leaks can occur because of pixel misalignments between the DMD mask and the camera's sensor; because of the binary rasterization of epipolar lines; and because of projector/camera defocus. To make indirect-only acquisition robust to such effects, we slightly dilate the "off" regions on the generated masks. This reduces the occurrence of such leaks at the expense of a slight reduction in light efficiency. We found this approach to be very effective in practice; a similar idea was used in [5].

Epipolar-only imaging We generate epipolar-only video by operating the camera at 56fps and configuring the DMD of our low-speed prototype as follows:

- *odd video frames:* display 48 all-on mask/pattern pairs
- *even video frames:* display a sequence of 48 indirect-only mask/pattern pairs.

Epipolar-only video at 28fps is generated by (1) scaling the odd frames by 0.25 to account for the reduced intensity of indirect-only imaging (Eq. (11)) and (2) subtracting in real time the even frames from the scaled odd ones.

Indirect-invariant imaging and indirect-invariant 3D reconstruction We generate a sequence of 800 mask/pattern pairs for each of 9 grayscale structured-light patterns, as illustrated in rows 2-4 of Figure 2. We then capture one raw image of the scene for each of the 9 generated mask/pattern sequences. These 9 images are supplied, unaltered, to the 3D reconstruction algorithm.

One-shot, multi-pattern, indirect-invariant imaging We generate a sequence of 792 mask/pattern pairs, as outlined in row 5 of Figure 2, and upload them to the DMDs. We then apply the algorithm outlined in Section 5 independently to each frame of the raw live video stream.

E. Discussion of Videos in `SLT-supp.zip`

E.1. Videos in directory `figure5_videos/`

These videos are meant to be used in conjunction with Figure 5 in the paper.

- *Stripe scan:* The purpose of this video is to show that when we sweep a stripe along the scene, the indirect light received at an epipolar line (red line in the video) can be very significant.
- *Spot scan:* In this video we restrict projector illumination to the corresponding epipolar line. This enforces epipolar-only transport for pixels on the epipolar line shown in red. The video shows that the indirect light received at the red epipolar line is minimal, even for a highly-complex scene. Taken together, the stripe-scan and spot-scan videos demonstrate that the bulk of indirect light reaching the red epipolar line originates from projector pixels *outside* the corresponding epipolar line.
- *Acquiring $\mathbf{T}_{ee}$:* This video shows how an epipolar block $\mathbf{T}_{ee}$ can be acquired: for every frame in the spot-scan video, we collect all pixels along the red epipolar line and place them as a column of $\mathbf{T}_{ee}$, also shown in red. The video shows how the epipolar block is acquired, column by column, by sliding the spot along the epipolar line from left to right.
- *Acquiring $\sum_f \mathbf{T}_{ef}$:* An analogous visualization of the acquisition of $\sum_f \mathbf{T}_{ef}$, *i.e.*, by sweeping a vertical stripe from left to right on the projector plane.

E.2. Videos in directory `live_SLT_imaging/`

We show conventional, indirect-only and epipolar-only streams for a variety of scenes. These were recorded live with the prototype shown in Figure 8 of the paper. All streams were captured with the same camera and the same settings (exposure time, white balance, *etc.*) The three streams shown in each video were captured sequentially, with the only difference being the masks/projection patterns used. This is because we only have one color SLT prototype and it can operate in one of three modes at any given time (*i.e.*, conventional, indirect-only, epipolar-only).

[1] We emphasize that flicker is a purely subjective sensation that may be experienced by users who view the scene directly, without the benefit of the DMD mask. In particular, flicker *does not* occur in the videos captured by our prototypes.

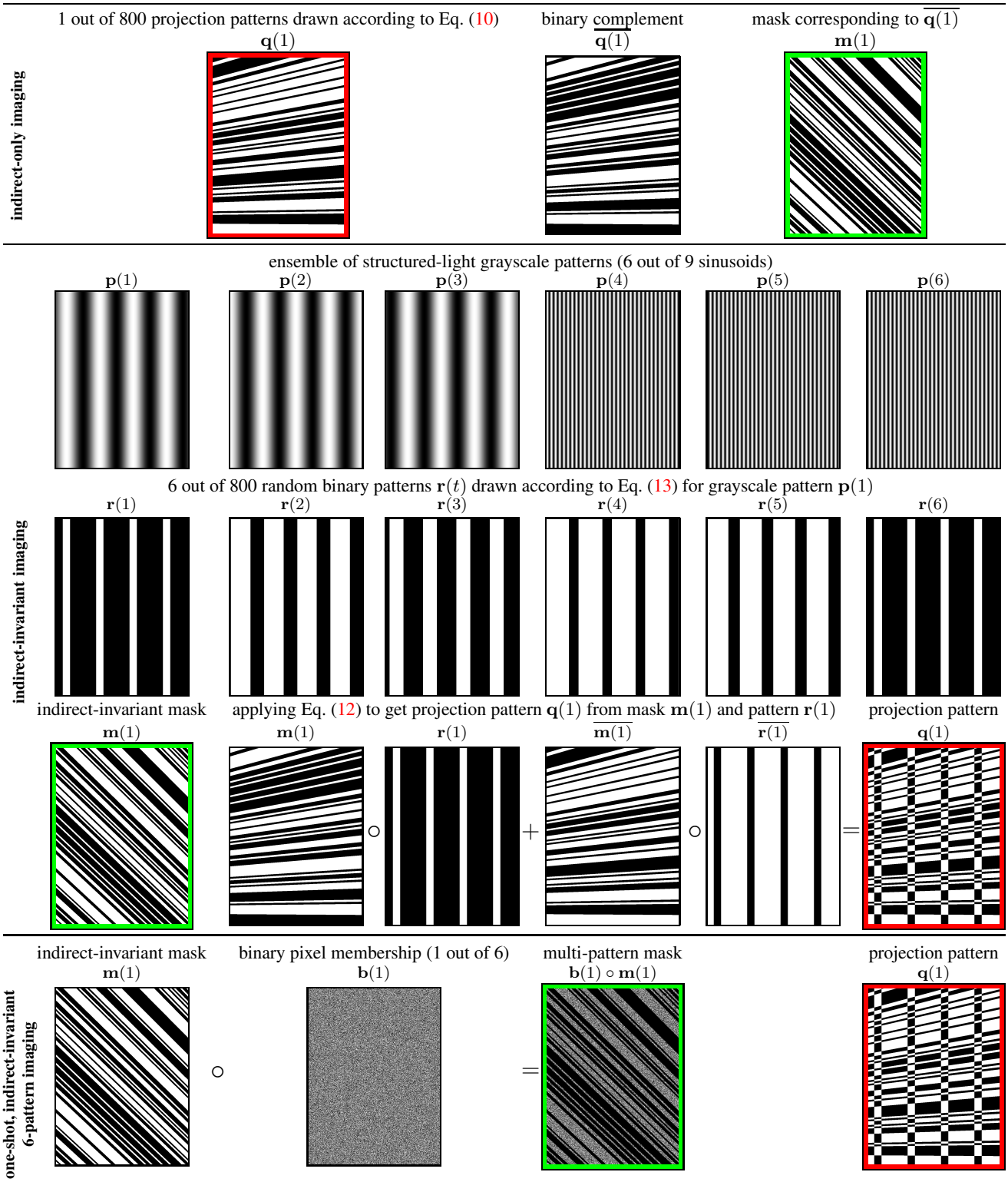

Figure 11: Deriving random pattern/mask pairs for three cases of SLT imaging. The derived patterns and masks are indicated with red and green borders, respectively. **Row 1:** For indirect-only imaging, the patterns and masks are constant along epipolar lines, with approximately half of them "on." **Row 2:** Six of the nine structured-light patterns we used. **Rows 3-4:** The masks for indirect-invariant imaging are identical to those for indirect-only imaging but the projection patterns differ. To generate them for a given grayscale structured-light pattern, we first generate a random sequence of binary patterns (Row 3) and then use that sequence, along with the sequence of masks, to compute the projection patterns. Row 4 shows one such example. **Row 5:** We generate pattern/mask pairs for 6-shot imaging as follows: (1) create 6 random binary images representing pixel membership for each pattern; (2) generate a sequence of 132 indirect-invariant binary pattern/mask pairs for each of 6 grayscale structured-light patterns, as outlined in Rows 2-4; (3) use the 792 projection patterns as is, and (4) multiply the masks element-wise with the associated pixel memberships. Row 5 shows one such calculation, for grayscale pattern $\mathbf{p}(1)$.

Conventional streams were captured under an all-on projection pattern with an all-on binary mask. Indirect-only streams were captured with 96 mask/pattern pairs, as explained in Section D. The conventional and indirect-only videos are recordings of the live, raw, video stream output by the camera.[2] Epipolar-only videos are created by pairwise differencing of adjacent video frames.

- *Candle:* A translucent candle.
- *Foam:* A piece of packing foam. The apparent speckles in the epipolar-only video are real: they correspond to momentary specular reflection from small, shiny membranes on the foam's surface.
- *Faux fur:* Note the marked difference between the epipolar-only component, which appears very shiny due direct near-specular reflection off the faux fur, versus the diffuse appearance of the indirect-only component, caused by sub-surface scattering. The occasional yellow tint in the epipolar-only component is due to saturation.
- *Glass:* This is the beer glass shown in Figures 5 and 1. Note that the glass appears essentially opaque in the epipolar-only component. This is because the light transmitted through the glass undergoes refraction, yielding non-linear paths that almost never lie on a single epipolar plane.
- *Hand:* Note the veins visible in the indirect-only component; also note the significant difference in apparent color of the hand in the indirect-only and epipolar-only components (due to sub-surface absorption and direct surface reflection, respectively).
- *Mug:* This is the mug shown in Figure 1. Note that artifacts appear occasionally on the white background behind the mug in the epipolar-only video. These occur because we do not compensate for motion when doing frame differencing. None of these artifacts appear in the indirect-only video, where no such differencing takes place.
- *Metal:* The indirect-only video clearly shows the very interesting caustics formed by the surface of this shiny metal plate. These caustics move very quickly; as a result, the frame-differencing we do for epipolar-only imaging causes ghosting on the white background. Again, none of these artifacts appear in the indirect-only video, which does not rely on frame differencing.
- *Water:* This example demonstrates our ability to successfully image a highly-complex, time-varying phenomenon, such as pouring water into a glass. As in the previous examples, the water appears mostly specular and opaque in the epipolar-only video whereas the caustics produced by light are clearly visible in the indirect-only video.
- *Wet hand:* The indirect-only video makes apparent the very dramatic changes in a hand's reflectance properties when water flows over it. We hypothesize that these changes are caused by scattering in the thin film of water flowing over the hand.
- *Small candle, paper:* More examples.

[2] To reduce file size for inclusion in the supplementary materials, it was necessary to compress these videos. As a result, some compression artifacts may be present.

E.3. Videos in directory `transport_robust_3D/`

This directory contains input images and 3D reconstruction results for the two scenes shown in rows 2-3 of Figure 9. Rows 2-4 of Figure 2 show the derivation of one of the mask/pattern pairs we used for this purpose.

- *Conventional phase shift input:* 9 input images acquired by conventional projection of phase-shifted patterns.
- *Conventional phase shift recontruction:* raw 3D points reconstructed from those input images.
- *Indirect-invariant phase shift input:* 9 input images acquired by indirect-invariant imaging with the same 9 patterns.
- *Indirect-invariant phase shift reconstruction:* raw 3D points reconstructed from those input images.
- *Conventional stripe-based recontruction:* as another example, we show reconstruction results obtained by sweeping a vertical stripe across the scene, as in conventional triangulation-based 3D laser scanning (768 images total). Despite the fact that stripe scanning relies on a much larger input dataset, our approach produces comparable results for the bowl scene and a far more complete model for the face scene.

E.4. Videos in directory `live_3D_capture/`

These videos provide details on the process of reconstructing a dynamic scene from video acquired by one-shot, 6-pattern indirect-invariant imaging. Row 5 of Figure 2 shows one of the mask/pattern pairs we constructed for this purpose.

- *Hand raw:* 169 frames of a moving hand, recorded live using one-shot, indirect-invariant, multi-pattern imaging.
- *Hand demosaic:* the 6 full-resolution videos resulting from demosaicing each video frame individually to obtain indirect-invariant views of the scene under 6 phase-shifted patterns.
- *Hand depth, albedo:* the raw, unprocessed depth and albedo maps reconstructed by applying conventional phase shifting to the demosaiced images.
- *Hand view albedo 1,2:* the reconstructed and texture-mapped geometry, shown from two different view-

points.

- *Hand view depth 1,2:* the reconstructed depth map, shown from two different viewpoints without texture mapping.

References

[1] V. Guillemin and A. Pollack. Differential Topology. *Prentice-Hall, 1974*.

[2] J. T. Kajiya. The Rendering Equation. *Proc. SIGGRAPH'86*.

[3] J. J. Koenderink. Solid Shape. *Cambridge University Press, 1990*.

[4] W. Rudin. Principles of Mathematical Analysis. *McGraw-Hill, 1976*.

[5] M. O'Toole, R. Raskar, K. N. Kutulakos. Primal-dual coding to probe light transport. *Proc. SIGGRAPH'12*.