Depth From Texture Integration

Mark Sheinin and Yoav Y. Schechner
Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology, Haifa, Israel

We present a new approach for active ranging, which can be compounded with traditional methods such as active depth from defocus or off-axis structured illumination. The object is illuminated by an active textured pattern having high spatial-frequency content. The illumination texture varies in time *while* the object undergoes a focal sweep. Consequently, in a single exposure, the illumination textures are encoded as a function of the object depth. Per-object depth, a particular illumination texture, with its high spatial frequency content, is focused; the other textures, projected when the system is defocused, are blurred. Analysis of the time-integrated image decodes the depth map. The plurality of projected and sensed color channels enhances the performance of the process, as we demonstrate experimentally. Using a wide aperture and only one or two readout frames, the method is particularly useful for imaging that requires high sensitivity to weak signals and high spatial resolution. Using a focal sweep during an exposure, the imaging has a wide dynamic depth range while being fast.

***Index Terms*—computational imaging, focal sweep, three-dimensional shape recovery, active illumination, electrically tunable lens**

I. Introduction

WIDE aperture imaging is beneficial in several respects: (a) increased light gathering capacity, for sensing dim objects at a good signal-to-noise ratio (SNR); (b) increased lateral resolution by narrowing the diffraction-limited point-spread function in focus; (c) enhanced axial resolution thanks to discrimination of in-focus vs. defocused blurred content [9], [33]. The latter discrimination and resolution increase with the numerical aperture. For these three reasons, wide aperture imaging is particularly suitable for microscopy [30], [31] and macro-imaging at close range. Due to the close range, active structured illumination is highly effective [26], increasing the reliability and resolution of depth estimation [1], [2]. Structured illumination patterns often have high spatial-frequency content, making focus/defocus discrimination highly effective.

A wide-aperture imager can significantly widen its operational depth dynamic range using a *focal sweep* [6], [19], [22], [37]. The camera's plane of focus is swept through the scene (or vice versa) during an exposure. As a result, all object points within the wide range of the sweep appear in focus at some time instance during the exposure, irrespective of their depth. Moreover, the entire sweep is time integrated and can be captured in a single frame. Hence, in a single focal-sweep image, the signal-to-readout-noise ratio is significantly better than in frames read out per focal step, each after a fraction of the exposure time [6], [14]. Thus focal sweep offers new benefits in addition to those listed above: (d) wide depth dynamic range, high speed, and increased SNR. However, the integrated exposure of a focal sweep loses benefit (c) above: the result is insensitive to depth.

Nevertheless, we suggest that through active illumination, wide aperture imaging can overcome the shortfall mentioned above, enabling simultaneously benefits (a,b,c,d). The main idea is that during a *single* focal sweep, the spatial illumination texture changes in time (see Fig. 1). In a single exposure, the illumination textures are encoded as a function of object depth. Per-object depth, a particular illumination texture, with its high-spatially content, is focused; the other textures, projected

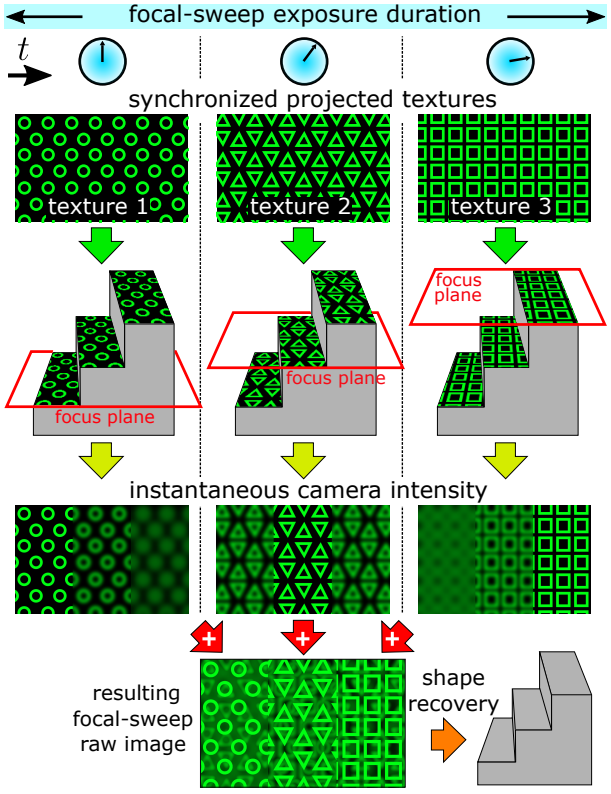

Fig. 1. Depth from texture integration. During image exposure, the illumination texture varies spatiotemporally. Focal-sweep is performed during this image exposure. A unique texture thus appears focused on the object surface, according to the the object distance. The time-integrated texture in the resulting image encodes object depth, enabling shape recovery.

Corresponding authors: M. Sheinin and Y. Schechner (emails: marksheinin@gmail.com, yoav@ee.technion.ac.il).

when the system is defocused, are blurred there. Analysis of the time-integrated image decodes the depth map.

The main principle of this work is time-integration of illumination textures during a simultaneous camera focal sweep. This principle can be applied in various camera-illumination configurations, including coaxial [32], confocal [35] and triangulation-based off-axis lighting. Off-axis lighting in microscopy can rely on speckle fields created by laser interference. Moreover, texture integration can work in conjunction with existing methods for depth estimation. These include projection of quasi-random dot patterns [15], volumetric stacking of lighting masks [16] and off-axis parallax of the projector relative to a camera [11].

II. Background: Focal Sweep Imaging

Denote the image coordinates by vector $\mathbf{x}$. Let $\tilde{I}^{\mathrm{AIF}}(\mathbf{x})$ denote the all-in-focus (AIF) intensity at $\mathbf{x}$. It would have been obtained had the camera had an infinite depth of field. The intensity $\tilde{I}^{\mathrm{AIF}}(\mathbf{x})$ has units of [graylevel/sec] and it expresses a projection of a domain illuminated by ambient light. Due to the camera's finite depth of field, only a narrow range is in focus. For a camera focused at distance u and an object of distance $d(\mathbf{x})$, the optical blur point-spread function is $k\left[\mathbf{x}|d(\mathbf{x}),u\right]$. Object surfaces lying near the focus plane u appear sharp, while object surfaces away from u are increasingly blurred according to $k\left[\mathbf{x}|d(\mathbf{x}),u\right]$.

Let $u(t)$ be the camera's focus plane at time t. In focal sweep imaging, $u(t)$ is swept over a range $[u_{\min},u_{\max}]$ during a *single* camera exposure lasting T_{exp} seconds (see Fig. 2). A focal-sweep image $I^{\mathrm{sweep}}(\mathbf{x})$ in [graylevel] units is given by

$$\begin{aligned} I^{\mathrm{sweep}}(\mathbf{x}) &= \int_0^{T_{\mathrm{exp}}} \int_{\mathbf{y}} \tilde{I}^{\mathrm{AIF}}(\mathbf{y}) k\left[(\mathbf{x}-\mathbf{y})|d(\mathbf{x}),u(t)\right] d\mathbf{y}dt \\ &= \int_0^{T_{\mathrm{exp}}} \tilde{I}^{\mathrm{AIF}}(\mathbf{x}) * k\left[\mathbf{x}|d(\mathbf{x}),u(t)\right] dt. \end{aligned} \tag{1}$$

At every $\mathbf{x}$, the resultant $I^{\mathrm{sweep}}(\mathbf{x})$ contains both sharp and defocused contributions of $\tilde{I}^{\mathrm{AIF}}(\mathbf{x})$. Prior works [14], [17], [19], [37] have shown that focal sweeps typically yield a point-spread function which is nearly depth-independent

$$\int_0^{T_{\mathrm{exp}}} k\left[\mathbf{x}|d(\mathbf{x}),u(t)\right] dt \approxeq K(\mathbf{x}). \tag{2}$$

Then,

$$I^{\mathrm{sweep}} \approxeq \tilde{I}^{\mathrm{AIF}}(\mathbf{x}) * K(\mathbf{x}). \tag{3}$$

Deconvolving $I^{\mathrm{sweep}}(\mathbf{x})$ with kernel $K(\mathbf{x})$ can thus retrieve $\tilde{I}^{\mathrm{AIF}}(\mathbf{x})$.

III. Imaging using Focal Sweep Codes

A. Approach Overview

Texture integration involves three *simultaneous* and synchronized processes. (a) A camera is exposed to the scene for T_{exp} seconds. (b) During the exposure, an electrically tunable lens (ETL) sweeps the focus plane, resulting in a focal-sweep image. (c) A projector synchronized with the ETL projects spatial patterns that temporally change during T_{exp}. Thus, the recorded image time-integrates the projected spatial patterns.

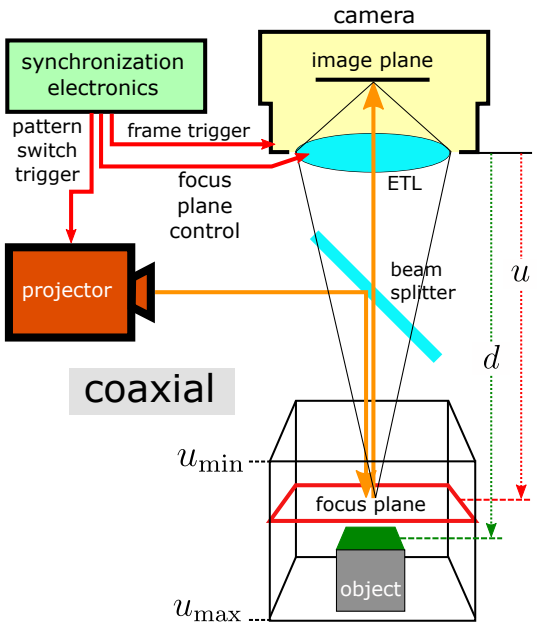

Fig. 2. A coaxial configuration. An electronically tunable lens (ETL) mounted on a camera creates a focal-sweep of a volume domain in a *single* exposure. Meanwhile, the domain is illuminated by a coaxial projector. During the sweep, the projector projects spatial textures that change temporally. The resulting single frame is depth-dependent. Additional configurations for texture integration are shown in Fig. 9.

The resulting single frame encodes depth-dependent textures (Section III-B). In this paper, we use 'texture' and 'pattern' interchangeably.

Several imaging configurations can exploit temporal integration of spatial textures. Here, we mainly focus on a *coaxial setting* (Fig. 2): a camera equipped with an ETL images the domain through a beam splitter, while a projector illuminates the domain coaxially. Additional optical configurations are discussed in Section VIII. In the setup we used, the projector's depth-of-field is much longer than that of the camera's optics, yielding negligible[1] projector defocus [8].

B. Image Formation Model

Texture integration relies on projecting a sequence of spatial high-frequency patterns during a single focal-sweep exposure. During the focal sweep, at each time t, hence focal position $u(t)$, the projector illuminates the object with a distinct spatial texture. Let $P_t(\mathbf{x})$ [graylevel/sec] denote the image irradiance at $\mathbf{x}$ corresponding to a *projector's* spatial pattern at time t. Henceforth, assume that ambient illumination is negligible with respect to the projector's illumination, namely that $\tilde{I}^{\mathrm{AIF}}(\mathbf{x}) \ll P(\mathbf{x})$. For the moment, texture projection uses only a single color band. Multi-spectral projection is discussed in Sec. VI.

In analogy to Eq. (1), due to projector illumination, the recorded image is

$$I(\mathbf{x}) = \int_0^{T_{\mathrm{exp}}} \rho(\mathbf{x}) P_t(\mathbf{x}) * k\left[\mathbf{x}|d(\mathbf{x}),u(t)\right] dt, \tag{4}$$

[1] An additional ETL fitted in a projector can control projector defocus and specifically remove the defocus, in synchrony with the camera-mounted ETL.

where $\rho(\mathbf{x})$ expresses the object albedo. A *discrete* focal sweep uses N focal steps denoted by u_n, where $n = 1, 2, .., N$. Each focal step remains still for T_{exp}/N seconds. Then, Eq. (4) becomes

$$I(\mathbf{x}) = \frac{T_{\text{exp}}}{N} \sum_{n=1}^{N} \rho(\mathbf{x}) P_n(\mathbf{x}) * k\left[\mathbf{x}|d(\mathbf{x}), u_n\right]. \tag{5}$$

C. Plane Response Function

Consider a uniform white planar object for which $\rho(\mathbf{x}) = 1$, positioned at distance d parallel to the camera's focus plane, as in Fig. 3. Then Eq. (4) yields

$$C(\mathbf{x}, d) \equiv \int_0^{T_{\text{exp}}} P_t(\mathbf{x}) * k\left[\mathbf{x}|d, u(t)\right] dt. \tag{6}$$

Observing Eq. (6), the texture in $C(\mathbf{x}, d)$ is a temporal integral of different spatial textures, each spatially filtered by a different optical blur according to d. In essence, image $C(\mathbf{x}, d)$ constitutes a *response function*: it is the texture integration's response to a uniform planar object (Fig. 3). Thus we refer to $C(\mathbf{x}, d)$ as a "plane-response function." Using a discrete focal sweep, Eq. (6) becomes

$$C(\mathbf{x}, d) = \frac{T_{\text{exp}}}{N} \sum_{n=1}^{N} P_n(\mathbf{x}) * k\left[\mathbf{x}|d, u_n\right]. \tag{7}$$

Suppose the response $C(\mathbf{x}, d)$ is known for any d. Then, from Eqs. (4)-(7), a texture-integrated image can be rendered for an arbitrary uniform object having depth $d(\mathbf{x})$ and constant albedo ρ, by

$$\hat{I}(\mathbf{x}) = \rho C\left[\mathbf{x}, d(\mathbf{x})\right]. \tag{8}$$

IV. Depth Recovery

A set of discrete depth steps $\mathcal{D} = \{d_m\}_{m=1}^{M}$ spans the axial domain. The set $\boldsymbol{C} = \{C(\mathbf{x}, d_m)\}_{d_m \in \mathcal{D}}$ denotes the corresponding axial samples of the plane-response function. When observing an object of interest, a single focal-sweep image $I(\mathbf{x})$ of the object is obtained following Eqs. (4,5). This frame becomes the input to a depth recovery procedure. Denote by $\mathcal{L}(\mathbf{x})$ a small image patch centered at $\mathbf{x}$. Let us match the content of $I(\mathbf{x})$ in patch $\mathcal{L}(\mathbf{x})$, to responses extracted from $C(\mathbf{x}, d_m)$ per m, in a corresponding spatial patch. The response $C(\mathbf{x}, d_m)$ that best matches $I(\mathbf{x})$ in $\mathcal{L}(\mathbf{x})$ indicates d_m as a primary candidate for the depth $d(\mathbf{x})$, as in Fig. 4.

Image noise, sharp albedo variation or large spatial gradients of $d(\mathbf{x})$ are inconsistent with the model of Section III-C. Hence, they may yield erroneous matches. Therefore, a prior on object shape is used. Let $d(\mathbf{x}) \in \mathcal{D}$. The depth-map is the set $\mathcal{R} = \{d(\mathbf{x})\}_{\forall \mathbf{x}}$. Define a cost function for a depth-map

$$E(\mathcal{R}) = \sum_{\mathbf{x}} D_{\mathbf{x}}\left[d(\mathbf{x})\right] + \lambda \sum_{\mathbf{x},\mathbf{x}'} V_{\mathbf{x},\mathbf{x}'}\left[d(\mathbf{x}), d(\mathbf{x}')\right]. \tag{9}$$

Here $V_{\mathbf{x},\mathbf{x}'}\left[d(\mathbf{x}), d(\mathbf{x}')\right]$ penalizes dissimilar depths at neighboring pixels $\mathbf{x}, \mathbf{x}'$ [3], essentially expressing a smoothness prior weighted by λ. The data term $D_{\mathbf{x}}[d(\mathbf{x})]$ penalizes texture mismatch between $I(\mathbf{x})$ and the response $C[\mathbf{x}, d(\mathbf{x})]$ in $\mathcal{L}(\mathbf{x})$.

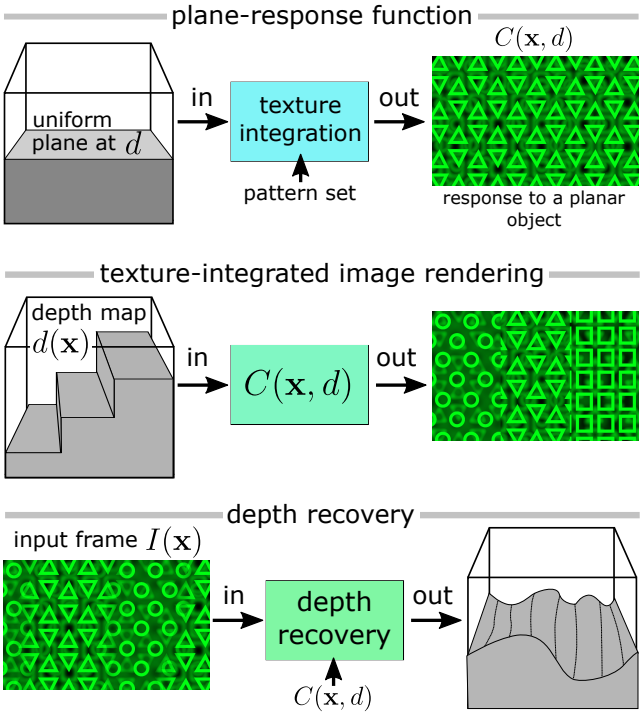

Fig. 3. **Top:** The plane-response function $C(\mathbf{x}, d)$ describes the response of texture integration to a uniform planar object at distance d. **Middle:** Response $C(\mathbf{x}, d)$ may render texture-integrated images of objects having a uniform albedo. **Bottom:** $C(\mathbf{x}, d)$ is used to solve the inverse problem: estimate object shape from a texture-integrated image.

To define $D_{\mathbf{x}}[d(\mathbf{x})]$, denote by $\text{ZNCC}^{\mathcal{L}}[\cdot, \cdot]$ the zero-normalized cross-correlation operator [5] in $\mathcal{L}(\mathbf{x})$. We set

$$D_{\mathbf{x}}[d_m] = 1 - \text{ZNCC}^{L}\left[I(\mathbf{x}), C(\mathbf{x}, d_m)\right]. \tag{10}$$

Recalling Eq. (8), note that Eq. (10) is invariant to the (generally unknown) albedo ρ of a uniform object. Moreover, if the true depth in $\mathbf{x}$ is d_m, the object is uniform, and there is no noise, then $D_{\mathbf{x}}[d_m] = 0$. We recover the depth-map of a general (non-flat) object by minimizing E with respect to $\mathcal{R}$

$$\hat{\mathcal{R}} = \arg\min_{\mathcal{R}} E(\mathcal{R}). \tag{11}$$

See Fig. 4 for an example. The experimental and parameter settings of all results shown in this paper are detailed in Sec. IX.

Measuring the Plane-Response Set

Depth recovery relies on availability of the response set $\boldsymbol{C}$. While this set may be derived using elaborate theoretical or simulated models, it is often simpler [16] to empirically sample $\boldsymbol{C}$. We measure $C(\mathbf{x}, d_m)$ directly by imaging a white planar object which is placed on a motorized stage. Using the moving stage, the object is shifted axially in increments of Δd, in the depth span $[u_{\min}, u_{\max}]$. Depths at which $C(\mathbf{x}, d)$ is sampled are $d_m = u_{\max} - m\Delta d$. Section VII discusses limitations of axial resolution.

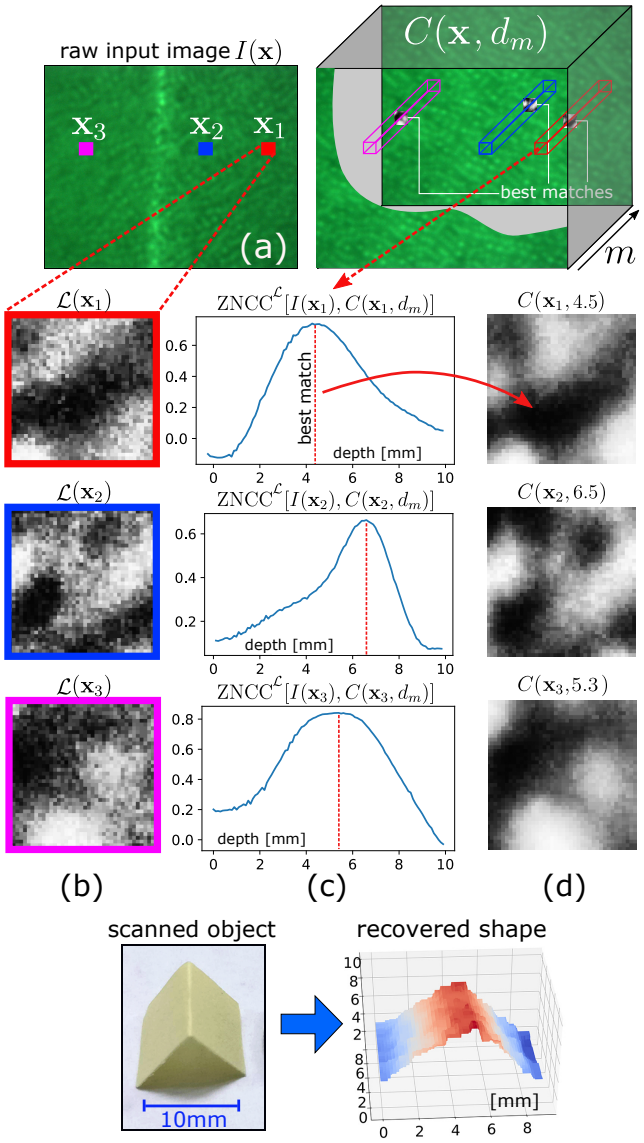

Fig. 4. **(a)** An input image $I(\mathbf{x})$ with three patches marked by red, blue and purple. **(b)** Zoom-in and contrast stretch of these patches. **(c)** Correlation between $I(\mathbf{x})$ and plane-response images, in each of the marked patches. **(d)** The response having the highest correlation indeed has visual similarity to the image patch. It indicates a prime candidate for object depth. The response here is shown in areas corresponding to the marked image patches.

V. The Set of Projected Textures

We experimented with several pattern sets for texture integration, guided by reconstruction quality and confusion matrices. Element $W[m, m']$ in confusion matrix $\mathbf{W}$ is

$$W[m, m'] = \mathbb{E}_{\mathbf{x}} \left\{ \mathrm{ZNCC}^{\mathcal{L}}[C(\mathbf{x}, d_m), C(\mathbf{x}, d_{m'})] \right\}, \quad (12)$$

where $\mathbb{E}_{\mathbf{x}}$ denotes spatial averaging over all pixels. A desirable confusion matrix $\mathbf{W}^{\mathrm{desired}}$ should have unit-values on the main diagonal and -1 values off the diagonal. The quality of $\mathbf{W}$ is assessed by how close it is to $\mathbf{W}^{\mathrm{desired}}$, in the sense of Frobenius norm $\|.\|_F$. We seek $\mathbf{W}$ for which

$$e(\mathbf{W}) = \|\mathbf{W} - \mathbf{W}^{\mathrm{desired}}\|_F^2 / M^2 \quad (13)$$

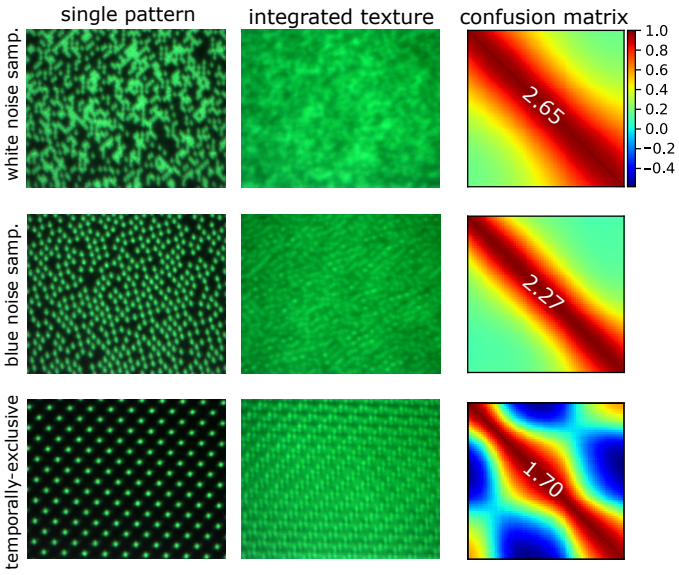

Fig. 5. **Left**: Sample spatial textures. **Middle**: The resulting integrated texture, when observing a white planar object. **Right**: The confusion matrix $\mathbf{W}$ per type of texture. Its corresponding $e(\mathbf{W})$ given in Eq. (13) is written on the diagonal.

is low.

We considered several binary patterns, e.g. rotated stripes [16]. We eventually settled on pseudo-random textures and temporally-exclusive codes. A pseudo-random spatial texture can be generated by white-noise sampling, i.e., a projector pixel is activated independently of other pixels (Fig. 5[Top]). However, in white noise, the energy in low spatial frequencies is just as significant as in high spatial frequencies. Low spatial frequencies yield broad regions having many active pixels. This leads to low contrast also in focus, hence degrading the useful signal.

Blue noise sampling [34] yields better textures than white noise. Here, an active pixel inhibits its neighbors from being active as well. This yields textures rich in energy at high-spatial frequencies (Fig. 5[Middle row]). An additional way to reach low values off the main diagonal of $\mathbf{W}$ is by temporally-exclusive codes. Here, each projector pixel $\mathbf{x}$ is active in only one of the projected textures, $n(\mathbf{x})$. This ensures that in $\mathbf{x}$, no focused texture illumination is repeated at $n' \neq n(\mathbf{x})$ (Fig. 5[Bottom]). However, as N increases, temporal-exclusivity reduces the amount of active pixels per focus step, thus degrading spatial resolution.

VI. Multi Spectral Patterns

So far, the model assumed projection using a single color channel. Let us generalize the discussion for multi-spectral projection of textures. Textures can be projected in infrared, as in Microsoft's Kinect, allowing infrared-based shape recovery, simultaneously with all-in-focus visible light imaging. However, to keep the discussion intuitive, we now discuss color channels, without loss of generality.

Let $\sigma \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$ denote the spectral band index. Now, during the focal sweep, different patterns are projected at each

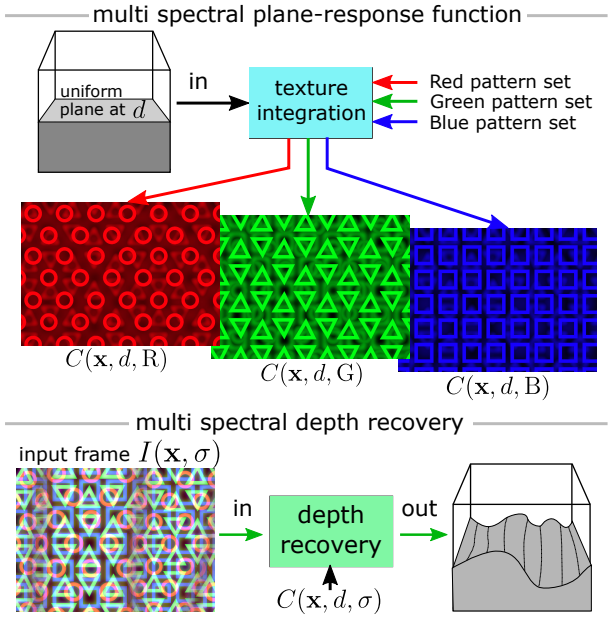

Fig. 6. Projected textures in multiple spectral bands. **Top:** Different textures are projected per band, resulting in a multi-spectral plane-response function. **Bottom:** Depth is recovered using multi-band data.

spectral band.[2] The plane-response function of Sec. III-C is thus expanded in the spectral dimension $C(\mathbf{x}, d) \rightarrow C(\mathbf{x}, d, \sigma)$ as illustrated in Fig. 6.

A. Multi-Spectral Illumination of an Object

One way in which depth sensing can fuse multiple bands, is to extract depth for each spectral band separately by inputting $I(\mathbf{x}, \sigma)$ in Eq. (10). The corresponding recovered depths per channel $\hat{d}(\mathbf{x}, \sigma)$ may then be fused using

$$\hat{d}(\mathbf{x}) = \frac{\sum_\sigma q(\mathbf{x}, \sigma) \hat{d}(\mathbf{x}, \sigma)}{\sum_\sigma q(\mathbf{x}, \sigma)}. \tag{14}$$

Here $q(x, \sigma)$ expresses per-band consistency of data with Eq. (8),

$$q(\mathbf{x}, \sigma) \equiv \max\left(0, \mathrm{ZNCC}^{\mathcal{L}}\left[I(\mathbf{x}, \sigma), \hat{I}(\mathbf{x}, \sigma)\right]\right). \tag{15}$$

A different use of multi-band imaging is normalization intended to decrease biases caused by spatial variations of the albedo $\rho(\mathbf{x})$. Let us first lay out the problem, by recalling Sec. IV. The data-fitting term relies on a model in which the albedo is uniform in each patch $\mathcal{L}(\mathbf{x})$. There, high-spatial frequency components are associated exclusively with the projected textures $P_n(\mathbf{x})$, while albedo has zero spatial frequency. Most real objects are not uniform, however. Low spatial frequency components in $\rho(\mathbf{x})$ may bias the estimation of d only slightly. On the other hand, high energy in high spatial frequency components of $\rho(\mathbf{x})$ may significantly bias the estimation results. It is desired that depth estimation would be less prone to spatial variations in $\rho(\mathbf{x})$. The core problem is that $I(\mathbf{x})$ is sensitive to these spatial variations. Countering the problem is possible by inputting to the algorithm of Sec. IV a representation whose sensitivity to spatial variations in $\rho(\mathbf{x})$ is lower than that of $I(\mathbf{x})$. This representation can be achieved using multi-band imaging, as we now describe.

Often there is strong correlation of albedo $\rho(\mathbf{x}, \sigma)$ between different spectral bands. Hence, $\rho(\mathbf{x}, \sigma)$ can be approximated by a superposition of albedo maps in all complementary channels, i.e., $\rho(\mathbf{x}, \sigma')$ where $\forall \sigma' \neq \sigma$. For example, channels $\sigma' \in \{\mathrm{R}, \mathrm{B}\}$ are complementary to the $\sigma = \mathrm{G}$ channel. There is generally no access to albedo maps in single-image acquisition. However, a color camera and a color projector enable traditional focal-sweep images (Eqs.1,3) at $\forall \sigma' \neq \sigma$, while texture integration is done at σ. Let us project spatial textures in one spectral band σ. Meanwhile, in the *complementary* spectral bands $\forall \sigma' \neq \sigma$, let the illumination be spatially *uniform*. Define

$$I^{\mathrm{multi}}(\mathbf{x}) = \frac{I(\mathbf{x}, \sigma)}{\sum_{\sigma' \neq \sigma} \alpha(\sigma') I(\mathbf{x}, \sigma')}, \tag{16}$$

where $\alpha(\sigma')$ are coefficients per channel. With proper selection of $\{\alpha(\sigma')\}_{\sigma'}$, the representation $I^{\mathrm{multi}}(\mathbf{x})$ can practically be rather insensitive to high spatial-frequency components of $\rho(\mathbf{x}, \sigma)$. Hence, in Eq. (10) instead of $I(\mathbf{x})$, it is beneficial to use $I^{\mathrm{multi}}(\mathbf{x})$.

The coefficients $\alpha(\sigma')$ are set using a least-squares approximation. For example, let $\sigma = \mathrm{G}$ and $\sigma' \in \{\mathrm{R}, \mathrm{B}\}$. Define matrix $\mathbf{A}$ and vector $\mathbf{b}$ by

$$\mathbf{A} \equiv \begin{bmatrix} I(\mathbf{x}_0, \mathrm{R}) & I(\mathbf{x}_0, \mathrm{B}) \\ I(\mathbf{x}_1, \mathrm{R}) & I(\mathbf{x}_1, \mathrm{B}) \\ I(\mathbf{x}_2, \mathrm{R}) & I(\mathbf{x}_2, \mathrm{B}) \\ \vdots & \vdots \end{bmatrix}, \quad \mathbf{b} \equiv \begin{bmatrix} I(\mathbf{x}_0, \mathrm{G}) \\ I(\mathbf{x}_1, \mathrm{G}) \\ I(\mathbf{x}_2, \mathrm{G}) \\ \vdots \end{bmatrix}. \tag{17}$$

Then, the coefficients are set by

$$\boldsymbol{\alpha} \equiv \begin{bmatrix} \alpha(\mathrm{R}) \\ \alpha(\mathrm{B}) \end{bmatrix} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b}, \tag{18}$$

where $\top$ denotes transposition.

B. Empirical Comparisons of Use of Color

An example of countering albedo variation using $I^{\mathrm{multi}}(\mathbf{x})$ is seen in Fig. 7. In addition, we tested color projection in three methods, as shown in Fig. 8:

(s1) Texture in a single spectral band, and no light in the other bands.

(s2) Multiple spectral bands, where a different random texture set is used per band.

(s3) Texture in a single band + uniform lighting in the complementary bands.

We found that method (s2) provided better lateral resolution, while method (s3) performed best in terms of noise and handling sharp albedo gradients.

VII. Limitations

This section discusses limitations regarding resolution and the dynamic range of the depth map. For a camera fitted with wide aperture optics (including an ETL), let the depth of field be 2δ. Around the object depth d, shifting the focus of a

[2]Often, the projector's spectral bands do not match those of the camera. This is handled by a procedure described in Appendix A.

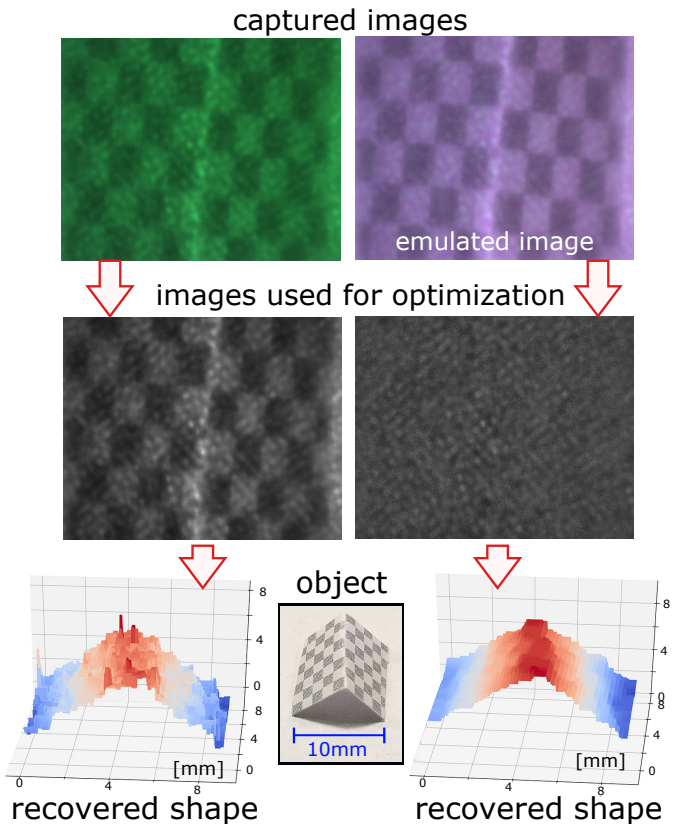

Fig. 7. **Left:** Only a single projector channel (green) illuminates the object. The illumination is textured. Results have errors due to albedo spatial edges. **Right:** Texture integration where the projected texture is green while the red and blue projector channels emit spatially uniform light. The red and blue channels normalize the green channel image (Eq. 16) as pre-processing, prior to depth estimation. This reduces the effect of albedo variations on the recovery.

camera in the range $[d-\delta, d+\delta]$ yields unnoticeable defocus blur. Hence, axial resolution is fundamentally limited by δ.

Recall that each spatial texture $P_t(\mathbf{x})$ corresponds to a focus (depth) setting $u(t)$. Hence, object depth d corresponds to $P_t(\mathbf{x})$ for which $|u(t)-d|$ is minimal, i.e., a texture projected when the ETL is focused nearest to d. Textures $P_t(\mathbf{x})$ for which $|u(t)-d| \gg \delta$ correspond to focus settings very far from d. Thus, these textures are too defocus-blurred to meaningfully affect the response $C(\mathbf{x}, d)$. For a strong response at any depth d, projected spatial textures should thus vary in time steps which correspond to depth increments Δu satisfying $\Delta u = \delta$.

The axial range $[u_{\min}, u_{\max}]$ is bounded as well. The bound is not only due to the finite dynamic range of the ETL, but also due to noise, as we show now. The number of patterns is

$$N \approx \frac{u_{\max}-u_{\min}}{\Delta u} \sim \frac{u_{\max}-u_{\min}}{\delta}. \tag{19}$$

Out of them, a handful of textures (N_{signal}) are projected at times t for which $|u(t)-d|$ is *not* much larger than δ. Hence, these few textures contribute a signal that, though slightly blurred, can still relate to the plane response function. The other textures are so defocus-blurred, that they essentially contribute a nearly uniform background radiance. This background contributes to *photon noise*. The variance of photon

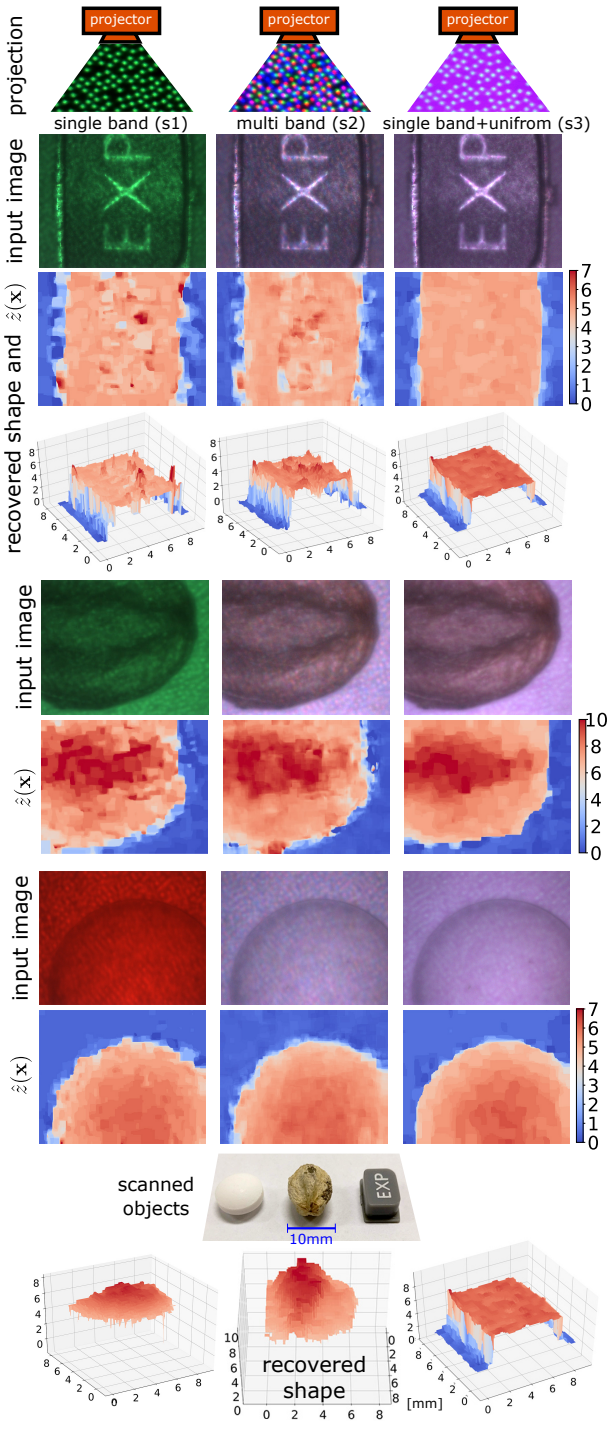

Fig. 8. Shape recovery using three types of projected sets. **Left:** Texture in a single spectral band, and no light in the other bands (s1). **Middle:** Multi band: all three projector colors project (different) random textures (s2). **Right:** Texture in a single band + uniform lighting in the complementary bands (s3). The bottom object (white pill) was scanned using the red LED in the single-band experiment.

noise is proportional to the overall integrated signal, including the background. Following Eq. (19), the photon-limited SNR is thus

$$\text{SNR} \approx \frac{N_{\text{signal}}}{\sqrt{N}} \sim \frac{N_{\text{signal}}\sqrt{\delta}}{\sqrt{u_{\max}-u_{\min}}}. \tag{20}$$

Attempting to the sense depth in a dynamic range $(u_{\max} - u_{\min})$ too broad may thus suffer from poor SNR.

Lateral resolution is limited by the projector and camera optics. The higher the projector spatial resolution, the smaller the correlation patch size $\mathcal{L}(\mathbf{x})$ can be. In the coaxial configuration (Fig. 2), projector defocus may be a limiting factor [8], [10]. Projector defocus is turned into an advantage [12] in a confocal setting (Fig. 9[Top]), where the projector is focused along with the camera. A speckle configuration (Fig. 9[Middle]) can provide diffraction-limited textures, whose feature size can reach down to a half optical wavelength.

VIII. Comparison to Related Methods

Depth from texture integration is most advantageous for microscopy of objects spanning a long depth-range. In microscopy, diffraction and SNR considerations essentially lead to an optical narrow depth of field. Here we discuss alternative methods and additional architectures. We empirically tested alternatives: depth from focus (DFF) and depth from defocus (DFD). Comparisons to texture integration are detailed in this section and in Sec. IX. The results appear in Fig. 10. In these tests, measures were taken for fair comparison in terms of total exposure duration and quality. This section also describes additional architectures for *depth from texture integration*, which are not described by Fig. 2.

DFF [13] acquires an N-frame sequence (focal stack) during a total acquisition time comparable to our T_{exp}. Thus, for dynamic objects, DFF requires very short exposures per individual frames in the stack. Short exposures may result in underexposed images that are unsuitable for depth recovery. Moreover, due to the need for repeated readout operations, N is limited by the camera readout speed. The exposure times per focal step within our single-frame texture integration (focal sweep) is $T_n = T_{\text{exp}}/N$. The exposure time per frame in the DFF stack is set as follows.

(a) $T_{\text{DFF}} = T_n$, while DFF relies on checkerboard pattern [21] structured illumination (Fig. 10a).

(b) $T_{\text{DFF}} = T_n$, while DFF uses uniform lighting (Fig. 10b). DFF results in both cases (a,b) were inferior to texture integration (Fig. 10e).

(c) $T_{\text{DFF}} = 3.5T_n$. Here DFF quality (Fig. 10c) was comparable to texture integration.

We used a software code of Ref. [25] for DFF.

DFD [7], [28], [29] requires only a pair of frames acquired during a time comparable to our T_{exp}. Hence the frames are well-exposed. Each frame is focused on a different distance u. DFD performance degrades when the object depth-range is much longer than the system's depth-of-field. In such cases, there may be object regions severely blurred simultaneously in both pair frames: blur there is indistinguishable, thus ill-conditioning DFD. As a preliminary test, we applied a DFD software code [7] on frame pairs from a well-exposed focal stack having 70 frames. Indeed, all the frame pairs had severely blurred regions, yielding large errors (Fig. 10d). In this preliminary experiment, texture integration appeared superior. However, refining the DFD analysis may potentially improve upon the results of Fig. 10d.

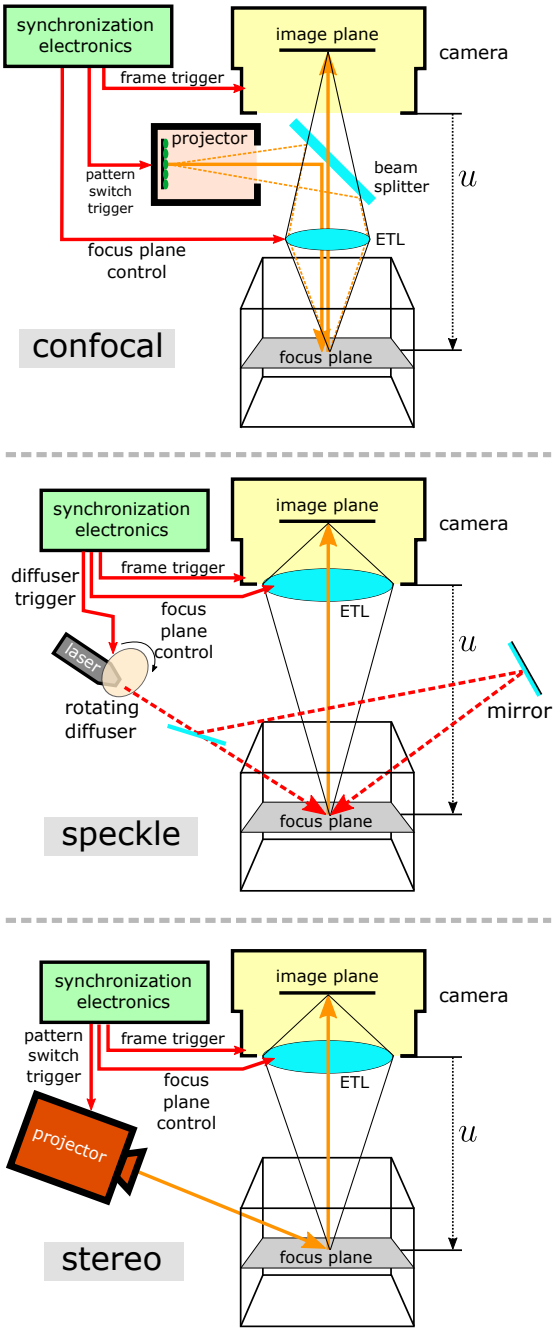

Fig. 9. Additional configurations for depth from texture integration. **Top:** Confocal configuration; both camera and projector are focused at the same plane, through the same electronically tunable lens. **Middle:** Speckle configuration; laser interference and diffraction create a spatial speckle pattern that can be varied in time in a repeatable manner. A way to implement this is by passing a laser through a rotating diffuser. **Bottom:** Stereo configuration; the projector is off-axis.

Triangulation: In general, in high-resolution microscopy, physical considerations inhibit the use of triangulation-based methods, such as Microsoft's Kinect [20]. Microscopy requires high numerical aperture optics to obtain the smallest features, which are limited by diffraction. A wide aperture yields a shallow depth of field, necessitating a focal-sweep (or stack). These considerations are incompatible with the assumptions

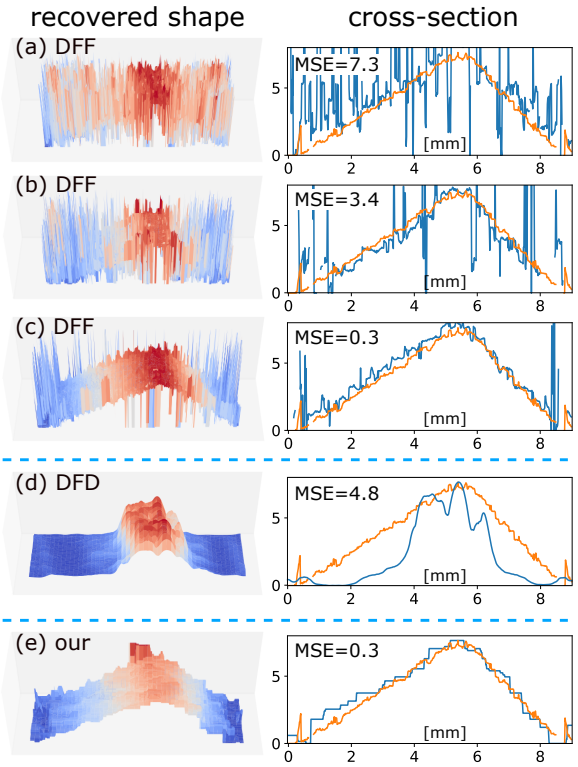

Fig. 10. Comparison to related methods. **Left**: Recovered shape. **Right**: Cross-sections of a recovery (blue) vs. the estimated ground truth (orange). **(a)** Structured light DFF. **(b)** Uniform light DFF. **(c)** Structured light DFF having 3.5 longer exposure time. **(d)** Two-frame DFD under uniform lighting. **(e)** Texture integration. Further details are in Section IX and Table I.

made by triangulation-based methods: geometrical optics, pinhole models, infinite depth-of-field.

In large scenes, where structured light for triangulation is suitable, depth is indicated by the lateral shift of an apparent pattern. While this shift can be very sensitive and provide high axial resolution, there can be large shadows and occlusions. Narrowing shadows and occlusions requires narrowing the camera-projector baseline to the limit of coaxial or confocal configurations (Figs. 2,9[Top]). In these configurations, depth sensitivity is largely based on depth-of-field, i.e., DFF, DFD and our texture-integration during sweep.

Texture-integration can be used in a structured-light triangulation configuration as well, as illustrated in Fig. 9[Bottom]. However, in this particular setting, we found no major advantage of integrating time-varying textures, relative to plain triangulation methods.

IX. Experimental Details

In this section we detail the experimental setup which yielded the results shown in Figs. 4,5,7,8,10. The setup is shown in Fig. 11. It includes an IDS UI-3240ML-C-HQ camera equipped with a 20mm extension tube, an Optotune EL-10-30 Ci ETL and a 50mm f2.8 Schneider lens. The object was placed $\approx$ 9cm from the camera. The focusing range was $u_{\max} - u_{\min} = 10$mm. Camera exposure was set to $T_{\exp} = 60$ms. During a single exposure, focal sweep was

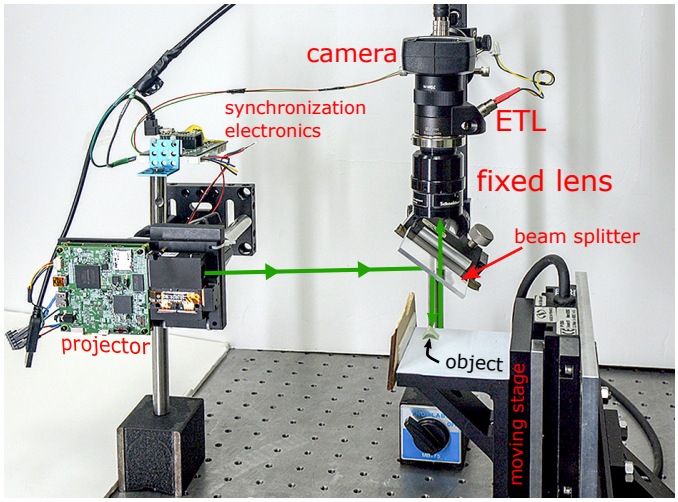

Fig. 11. Experimental setup. A camera images the scene through a beam splitter. The camera is equipped with an ETL that shifts the focus plane by approximately 10mm. A projector illuminates the object coaxially. A motorized stage is used to sample the plane-response function.

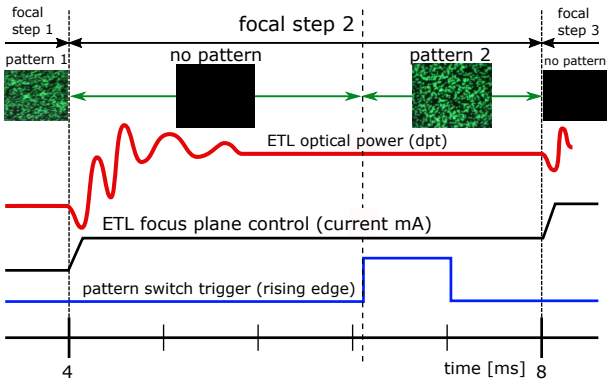

Fig. 12. Projection synchronization. At each focal step, the ETL control current is set for the desired optical power (diopters). Then, after a settling period, the spatial texture corresponding to that focal step is triggered.

realized in $N = 15$ discrete depth steps. To measure the plane-response function, a Newport VP-25XA stage powered by a ESP-300 driver axially shifted a planar object. The projector is a TI DLP3000 digital micromirror device (DMD) having 684×608 pixels. The ETL, projector and camera were synchronized using a custom electrical controller [18]. Being co-axial, there is no need to calibrate the camera-projector extrinsic geometry.

The total energy projected was controlled by limiting projection time per focal step (see Fig. 12). The projector is based on LED illumination: different color bands are created by respective LED colors. Simultaneous multi-LED projection on our specific hardware was too involved. We thus emulated methods (s2) and (s3) by capturing the scene under each LED separately, then summed the resulting raw Bayer images. After summations, grayscale levels over 255 were clipped to 255. Exposure settings relating to Fig. 10 are listed in Table I.

'Ground truth' depth was estimated by applying DFF on

TABLE I
SETTINGS FOR THE EXPERIMENT OF FIG. 10. THE TOTAL EXPOSURE PER STEP COMPRISES THE EXPOSURE TIMES OF THE RED, GREEN AND BLUE LIGHTS.

Method	Focal steps	Total exposure per step [us]	Red exp. per step [us]	Green exp. per step [us]	Blue exp. per step [us]	Projected content	MSE
(a) Depth from focus	15 frames	3080	440	2200	440	G-checkerboard pattern, R+B spatially uniform	7.3
(b) Depth from focus	15 frames	3080	440	2200	440	G+R+B spatially uniform	3.4
(c) Depth from focus	15 frames	10800	3600	3600	3600	G-checkerboard pattern, R+B spatially uniform	0.3
(d) Depth from defocus	2 frames	32400	3600x3	3600x3	3600x3	G+R+B spatially uniform	4.8
(e) Texture integration	15 steps	3080	440	2200	440	G-textures, R+B spatially uniform	0.3
Ground truth estimate	70 frames	32400	3600x3	3600x3	3600x3	G+R+B spatially uniform	

a 70-frame focal stack, in which each frame accumulated 32400us exposure time. The estimated 'ground truth' depth exhibited artifacts at the object boundaries, which where ignored when computing the MSE in Fig. 10. We experimented with two axial resolutions when sampling the response $\boldsymbol{C}$: 100 steps having $\Delta d = 0.1$mm in the experiments corresponding to Figs. 4,7 and 68 steps having $\Delta d = 0.15$mm in the experiments corresponding to Figs. 5,8,10. The patch $\mathcal{L}(\mathbf{x})$ is of size 41×41 pixels. In Eq. (9) we set $V_{\mathbf{x},\mathbf{x}'} = [d(\mathbf{x}) - d(\mathbf{x}')]^2 / \Delta d^2$ following [4]. Optimization (9) was run using [23] by first setting $\lambda = 0$. Then, Eq. (9) is computed again using $\lambda = 0.2$.

X. DISCUSSION

We present a novel imaging concept for fast sensing and recovery of depth in wide aperture settings. Texture integration can be useful in several imaging configurations. Similarly to other depth sensing methods, object specularities and subsurface scattering may degrade performance. Hence approaches to reduce these effects may need to be developed in the context of texture integration. Moreover, we believe that the projected textures can be systematically optimized to yield better performance.

The optimization in Section IV extracts discrete depth. However, the principle of depth from texture integration is not limited to this estimation algorithm. Continuous-valued depth maps can be estimated by using continuous optimization instead.

APPENDIX A

As mentioned in Section VI, camera and projector spectral bands often do not match. This creates crosstalk between spectral channels. An unmixing pre-process lowers this crosstalk.

Denote by $\{\mathrm{R}^{\mathrm{cam}}, \mathrm{G}^{\mathrm{cam}}, \mathrm{B}^{\mathrm{cam}}\}$ and $\{\mathrm{R}^{\mathrm{proj}}, \mathrm{G}^{\mathrm{proj}}, \mathrm{B}^{\mathrm{proj}}\}$ the camera and projector spectral bands, respectively. The crosstalk is modeled by

$$\begin{bmatrix} I(\mathbf{x}, \mathrm{R}^{\mathrm{cam}}) \\ I(\mathbf{x}, \mathrm{G}^{\mathrm{cam}}) \\ I(\mathbf{x}, \mathrm{B}^{\mathrm{cam}}) \end{bmatrix} = \mathbf{H} \begin{bmatrix} I(\mathbf{x}, \mathrm{R}^{\mathrm{proj}}) \\ I(\mathbf{x}, \mathrm{G}^{\mathrm{proj}}) \\ I(\mathbf{x}, \mathrm{B}^{\mathrm{proj}}) \end{bmatrix}, \qquad (21)$$

where $\mathbf{H}$ is a 3×3 color mixing matrix. Matrix $\mathbf{H}$ is calibrated by imaging a white object using the camera, while sequentially irradiating the object by a single projector spectral channel.

Then, let us image an arbitrary object using our system. Per pixel $\mathbf{x}$, the measured vector is $\mathbf{i}(\mathbf{x}) = [I(\mathbf{x}, \mathrm{R}^{\mathrm{cam}})\ I(\mathbf{x}, \mathrm{G}^{\mathrm{cam}})\ I(\mathbf{x}, \mathrm{B}^{\mathrm{cam}})]^\top$, where $\top$ denotes transposition. Spectral unmixing in this pixel is done by $\mathbf{H}^{-1}\mathbf{i}(\mathbf{x})$.

ACKNOWLEDGMENTS

We thank Tali Treibitz, Aviad Anvi and Judith Fischer for help with the project, Anat Levin, Guy Gilboa and Boris Spektor for useful discussions, and Paolo Favaro for providing us with code for testing. Y. Schechner is a Landau Fellow supported by the Taub Foundation. His work in this project is supported by the Israel Science Foundation (Grant 542/16). The research was partly carried in the Ollendorff Minerva Center. Minerva is funded through the BMBF.

REFERENCES

[1] S. Achar, and S. G. Narasimhan, Multi focus structured light for recovering scene shape and global illumination. In *Proc. ECCV*, 205-219, 2014.

[2] M. J. Amin, and N. A. Riza, Active depth from defocus system using coherent illumination and a no moving parts camera. In *Optics Comm.* 359, 135-145, 2016.

[3] Y. Boykov, O. Veksler, and R. Zabih, Fast approximate energy minimization via graph cuts. In *IEEE TPAMI*, 23(11), 1222-1239, 2001.

[4] Y. Boykov and V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *IEEE TPAMI*, 2004.

[5] K. Briechle, and U. D. Hanebeck, Template matching using fast normalized cross correlation. In *Proc. SPIE Optical Pattern Recognition XII*, 95-103, 2001.

[6] O. Cossairt, and S. K. Nayar, Spectral focal sweep: Extended depth of field from chromatic aberrations. In *Proc. IEEE ICCP*, 1-8, 2010.

[7] P. Favaro and S. Soatto, *3-D Shape Estimation and Image Restoration: Exploiting Defocus and Motion-Blur. Springer*, 2006.

[8] M. Gupta, Y. Tian, S. Narasimhan, and L. Zhang, (De) focusing on global light transport for active scene recovery. In *Proc. IEEE CVPR*, 2969-2976, 2009.

[9] Q. Guo, E. Alexander and, T. E. Zickler, Focal track: depth and accommodation with oscillating lens deformation. In *Proc. IEEE ICCV,* 966-974, 2017.

[10] D. Iwai, S. Mihara, and K. Sato, Extended depth-of-field projector by fast focal sweep projection. In IEEE *Tran. on Vis. and Comp. Grap.,* 21(4), 462-470, 2015.

[11] H. Kawasaki, S. Ono, Y. Horita, Y. Shiba, R. Furukawa, and S. Hiura, Active one-shot scan for wide depth range using a light field projector based on coded aperture. In *Proc. IEEE ICCV*, 3568-3576. 2015.

[12] H. Kawasaki, Y. Horita, H. Masuyama, S. Ono, M. Kimura and, Y. Takane, Optimized aperture for estimating depth from projector's defocus. In *Proc. IEEE 3DV*, 135-142, 2013.

[13] S. Kuthirummal, H. Nagahara, C. Zhou, and S. K. Nayar, Flexible depth of field photography. In *IEEE TPAMI*, 33(1):58-71, 2011.

[14] X. Lin, J. Suo, G. Wetzstein, Q. Dai and R. Raskar, Coded focal stack photography. In *Proc. IEEE ICCP*, 2013.

[15] A. Ma, and A. Wong. An inverse problem approach to computational active depth from defocus, In IOP *J. Phys.: Conf. Ser.* 1047(1), 12009, 2018.

[16] H. Masuyama, H. Kawasaki, and R. Furukawa, Depth from projector's defocus based on multiple focus pattern projection. In *IPSJ Trans. on Comp. Vision and App.*, 6, 88-92, 2014.
[17] D. Miau, O. Cossairt, and S. K. Nayar, Focal sweep videography with deformable optics. In *Proc. IEEE ICCP*, 2013.
[18] A. D. Mullen, T. Treibitz, P. L. Roberts, E. L. Kelly, R. Horwitz, J .E. Smith, and J .S. Jaffe, Underwater microscopy for in situ studies of benthic ecosystems. In *Nature Comm.*, 7, 2016.
[19] H. Nagahara, S. Kuthirummal, C. Zhou, and S. K. Nayar, Flexible depth of field photography. In *Proc. ECCV*, 60-73, 2008.
[20] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, KinectFusion: Real-time dense surface mapping and tracking. In *IEEE Int. Symp. on Mixed and Aug. Real.*, 127-136, 2011.
[21] M. Noguchi and S.K. Nayar, Microscopic shape from focus using active illumination. In *IEEE Proc. ICPR*, (1):147-152, 1994.
[22] Y. Peng, X. Dun, Q. Sun, F. Heide, and W. Heidrich, Focal sweep imaging with multi-focal diffractive optics. In *Proc. IEEE ICCP*, 2018.
[23] https://github.com/pmneila/PyMaxflow
[24] N. A. Riza, and S. A. Reza, Smart agile lens remote optical sensor for three-dimensional object shape measurements. In *Applied Optics*, 49(7), 1139-1150, 2010.
[25] S. Pertuz, Shape from focus. (https://www.mathworks.com/matlabcentral/fileexchange/55103-shape-from-focus), MATLAB Central File Exchange. 2018.
[26] G. Satat, M. Tancik, and R. Raskar, Lensless imaging with compressive ultrafast sensing. In *IEEE Trans. on Comp. Imag.*, 3(3):398-407, 2017.
[27] Y. Y. Schechner, and N. Kiryati, The optimal axial interval in estimating depth from defocus. In *Proc. IEEE ICCV*, (2):843-848, 1999.
[28] Y. Y. Schechner, and N. Kiryati, Depth from defocus vs. stereo: How different really are they? In *Int. J. Comp. Vision*, 39(2):141-162, 2000.
[29] Y. Y. Schechner and S. K. Nayar, Multidimensional fusion by image mosaics, In *Image Fusion: Alg. and App., Academic Press*, 193-221, 2008.
[30] W. J. Shain, N. A. Vickers, B. B. Goldberg, T. Bifano and, J. Mertz, Extended depth-of-field microscopy with a high-speed deformable mirror. In *Optics Letters*, 42(5), 995-998, 2017.
[31] W. J. Shain, N. A. Vickers, J. Li, X. Han, T. Bifano, and J. Mertz, Axial localization with modulated-illumination extended-depth-of-field microscopy. In *Biomedical Optics Express*, 9(4), 1771-1782, 2018.
[32] K. Tanaka, Y. Mukaigawa, H. Kubo, Y. Matsushita, and Y. Yagi, Recovering inner slices of translucent objects by multi-frequency illumination. In *Proc. IEEE CVPR*, 5464-5472, 2015.
[33] H. Tang, S. Cohen, B. L. Price, S. Schiller, and K. N. Kutulakos, Depth from Defocus in the Wild. In *Proc. IEEE CVPR*, 4773-4781, 2017.
[34] R. A. Ulichney, Dithering with blue noise. In *Proceedings of The IEEE*, 76(1):56-79 , 1988.
[35] S. Xiao, H. A. Tseng, H. Gritton, X. Han and, J. Mertz, Video-rate volumetric neuronal imaging using 3D targeted illumination. In *Scientific Reports*, 8(1), 7921, 2018.
[36] L. Xiao, F. Heide, M. O'Toole, A. Kolb, M. B. Hullin, K. N. Kutulakos, and W. Heidrich, Defocus deblurring and superresolution for time-of-flight depth cameras. In *Proc. IEEE CVPR*, 2376-2384, 2015.
[37] C. Zhou, D. Miau, and S .K. Nayar, Focal sweep camera for space-time refocusing. *Technical Report, Department of Computer Science*, 2012.