FreeCam3D: Snapshot Structured Light 3D with Freely-Moving Cameras

Yicheng Wu[1], Vivek Boominathan[1], Xuan Zhao[1], Jacob T. Robinson[1], Hiroshi Kawasaki[2], Aswin Sankaranarayanan[3], and Ashok Veeraraghavan[1]

[1] Rice University, Houston TX, USA
{yicheng.wu, vivekb, xz61, jtrobinson, vashok}@rice.edu
[2] Kyushu University, Fukuoka, Japan
kawasaki@ait.kyushu-u.ac.jp
[3] Carnegie Mellon University, Pittsburgh PA, USA
saswin@andrew.cmu.edu

Abstract. A 3D imaging and mapping system that can handle both multiple-viewers and dynamic-objects is attractive for many applications. We propose a freeform structured light system that does not rigidly constrain camera(s) to the projector. By introducing an optimized phase-coded aperture in the projector, we transform the projector pattern to encode depth in its defocus robustly; this allows a camera to estimate depth, in projector coordinates, using local information. Additionally, we project a Kronecker-multiplexed pattern that provides global context to establish correspondence between camera and projector pixels. Together with aperture coding and projected pattern, the projector offers a unique 3D labeling for every location of the scene. The projected pattern can be observed in part or full by any camera, to reconstruct both the 3D map of the scene and the camera pose in the projector coordinates. This system is optimized using a fully differentiable rendering model and a CNN-based reconstruction. We build a prototype and demonstrate high-quality 3D reconstruction with an unconstrained camera, for both dynamic scenes and multi-camera systems.

Keywords: Computational Photography; 3D Reconstruction; Coded Aperture; Structured Light

1 Introduction

3D scanning is one of the core technologies in many systems. For many upcoming applications, a depth map of the scene in the camera's viewpoint is not sufficient and it is equally important to localize the camera in a world-coordinate system. This problem gets all the more important when we have multiple cameras roaming in a shared space, as is the case in augmented reality, free-viewpoint videos, and indoor localization applications.

This paper provides an approach to obtain depth maps and localize one or many cameras, operating in a shared space, in a world coordinate system.

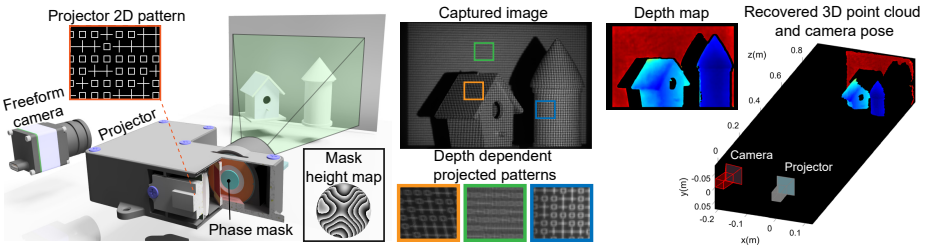

Fig. 1. Overview. (Left) Illustration of our system. An optimized phase mask is placed on the aperture of the projector to generate depth-dependent blur. The 2D pattern provides unique spatial features. (Center) Experimentally captured single-shot image by a freeform camera and the regions showing projected patterns at different 3D locations. (Right) Depth map and the camera (red) pose recovered with respect to the projector (gray) coordinates. Our system allows for multiple unconstrained participants/cameras to interact within the common world coordinate.

Our technique relies on a structured light system with a static projector that is decoupled from the camera(s); this projector, hence, provides a fixed (world) coordinate system for the scene against which cameras localize themselves. The projector displays a single static pattern, which is observed in part or full by any camera in the scene. Each camera decodes this image and localizes itself in the world coordinate system and, further, estimates a 3D map of the scene in its field of view. Since this is achieved *with a single image*, we enable a novel framework for single-shot self-localization and 3D estimation.

The advances made in this paper rely on three key ideas. First, to permit depth estimation without relying on triangulation, we use a projector that induces a depth-dependent defocus blur on the pattern projected on the scene. To further improve our ability to decode depth from the defocus blurs, we use an optimized phase mask on the pupil plane of the projection optics. Second, we design a projector pattern to help solve the correspondence problem between the projected pattern and the imaged pattern, especially in the presence of the defocus blur. The designed pattern is a Kronecker product between a random binary image, that provides global context, along with a textured local pattern that allows for local depth estimation via defocus. Third, we use a learning-based formulate that takes in the input image and predicts the X/Y correspondence as well as depth in the world coordinate system. The camera pose is estimated from this depth map using Perspective-n-Point (PnP) algorithms.

The proposed technique offers numerous advantages against traditional structured light and SLAM techniques. First, we can handle dynamic scenes since, at any time instant, only a single captured image is used for 3D estimation and self-localization. Second, the estimated 3D scan is in the world coordinate system as defined by the projector; this allows multiple cameras to share the same space seamlessly — a feature that is unique to our approach. Third, unlike structured light where the relative geometry between the camera and projector is known,

our technique is uncalibrated and estimates the camera's extrinsic parameters with respect to the projector automatically.

We summarize our contributions as follows:

- We propose a novel system for single-shot 3D reconstruction that relies on a fixed projector and freely-moving camera(s).
- Our system relies on an optimized phase mask in the projection optics. To perform this optimization, we build a fully differential model that contains the physical rendering (e.g., depth-dependent blurring and image warping) for end-to-end training, where the goal is to decode the image acquired by a camera. This simulation pipeline is directly applied to real experimental data without any finetuning.
- We build a prototype and demonstrate compelling 3D imaging performance using our prototype.

It is worth mentioning that, like other SL techniques, scene textures can reduce the performance of our method by corrupting the projected pattern. This can be reduced by operating in near-infrared wavelengths, similar to the Kinect system, as well as training our models with textured scenes.

2 Related work

2.1 Active depth sensing techniques

Active methods recover depth information by illuminating the scene with a coded light signal. Here, we provide three examples.

Time-of-flight (ToF). ToF cameras measure the depth based on the round trip time of a modulated light signal reflecting from the object [29]. While ToF cameras do provide single-shot depth estimates with little post-processing, both LIDAR and correlation-based approaches requir a strong coupling between the sensor and the active illuminant. When operating in a shared space, the devices tend to interfere with each other which causes artifacts in their reconstructions [19]. Further, the estimated depth maps are typically in a local coordinate system, which is not desirable for many applications.

Structured light (SL). SL is a triangulation-based method. The correspondence can be obtained by temporal coding or spatial coding [34]. Temporal coding methods are superior in spatial resolutions, but not suitable for dynamic scenes. For spatial coding, researchers have explored to recover depth from a snapshot based on the color [1, 23, 44, 38] or geometry [16, 40, 28] of the projected patterns. A recent class of techniques aims to enable 3D scanning using smartphones; since SL systems are usually fixed and static, whereas smartphones are mobile, there is a need for self-calibration. However, current approaches in this space either require additional information about the scene [7], or heavy computational cost for bundle adjustment [6].

Projection-based depth from defocus (DfD). There have been many approaches that use DfD using projectors [9, 30, 42, 5, 12]; a key advantage of such techniques is that we do not need to estimate correspondences. However, for a traditional lens-based system, the encoding of depth in defocus is not robust. This problem can be addressed by introducing a coded aperture in the projection optics [22, 41, 18, 17, 14]. These methods prevent the significant loss of information during defocus, as well as making it possible to decompose the overlapping pattern to obtain higher density and precision. It is worth pointing out that while our hardware is similar to those DfD systems, our novel algorithm allows the camera to be *unconstrained* while a standard DfD system requires the camera to be pre-calibrated and fixed.

2.2 Indoor localization

The goal of indoor localization is to obtain a device or user location in an indoor setting or environment. For a vision system, the camera pose consists of 6 degrees-of-freedom (DOF). A standard way to estimate camera pose is based on PnP algorithms [2, 21], which rely on a set of 3D points in the world coordinate and their corresponding 2D locations in the image. However, requiring known 3D points in world coordinates an unreasonable burden in many applications.

On the other hand, SLAM aims for estimating a map of an unknown environment while simultaneously keeping track of the location of the sensor. One key assumption is that the environment remains static when multiple frames are captured from the sensor. It means that SLAM has difficulty handling dynamic scenes. In comparison, our proposed method only requires a single image to recover the 3D environment as well as the camera pose.

2.3 Deep Optics

Recently, researchers have integrated deep learning algorithms to optimize computational imaging systems. The key idea is to treat the optical system as the first layer of a deep neural network. During the training, the free parameters of the optics as well as the deep networks are optimized end-to-end. This concept, often termed "deep optics", has found applications in demosaicing [3], depth estimation [43, 4, 13], extended depth of field [36], and high dynamic range [27, 37]. Our work follows the same spirit to optimize the phase mask design as well as the neural network.

3 Forward model

The goal of this section is to derive a differentiable physical model to simulate the captured camera image for any 3D scene and camera pose. As shown in Fig. 2, there are three main steps in the forward model: generating the 2D pattern that is projected, rendering the image in the projector's viewpoint with its depth-dependent defocus blur, and warping the pattern to create the captured image from the camera view.

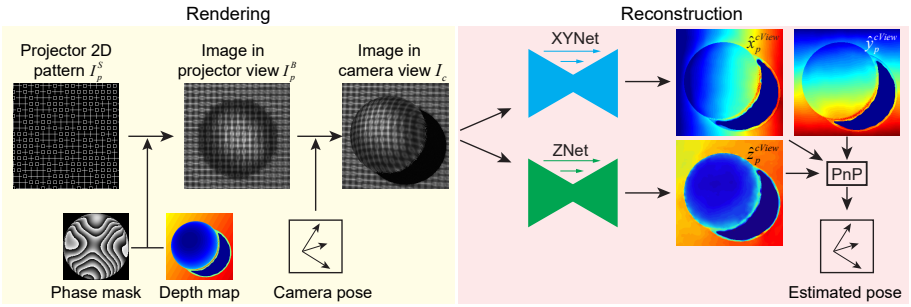

Fig. 2. System pipeline. (Left) The forward rendering part builds a physics-based model to simulate the captured camera image for any 3D scene and camera pose. (Right) From the single-shot image I_c, we first predict the 3D location in the projector coordinate. We then estimate the camera pose with a PnP solver. The pipeline is fully-differentiable, and can be trained end-to-end.

3.1 Projector 2D pattern design

There are two requirements for the projected pattern. First, to enable lateral (or x/y) localization, the pattern should contain unique local textures. Second, to enable axial (or z) localization, the pattern contains rich local textures to facilitate decoding of the defocus blur.

We propose to generate the pattern from a Kronecker product $\otimes$ between a global pattern I^{global} and two local patterns I_1^{local} and I_2^{local}. The final projected pattern I_p^S can be represented as

$$I_p^S = I^{\text{global}} \otimes I_1^{\text{local}} + (1 - I^{\text{global}}) \otimes I_2^{\text{local}} \tag{1}$$

We set I^{global} as a random binary pattern. I_1^{local} and I_2^{local} are cross and square, respectively. As we see from Fig. 2, the overall pattern still preserves a grid structure, which can be a useful clue for the reconstruction algorithm to estimate depth from the distorted image in the camera view.

3.2 Depth encoding with the phase mask

For a conventional lens-based SL system, the working depth range is limited by the depth of field, because the pattern has to be sharp for stereo matching algorithms. Instead, we estimate depth based on the defocus effect. Thus, the working depth range is increased significantly. To amplify the defocus effect for higher depth estimation, we insert a phase mask on the aperture plane so that the point spread function (PSF) varies rapidly over depth while the PSF size remains small. This approach follows a rich body of literature that improves depth resolution using specialized phase masks [31, 35, 43].

Point spread functions. For an incoherent system, the PSF is the squared magnitude of the Fourier transform of the pupil function [11].

$$PSF(h,z) = \left|\mathcal{F}\{A\exp[\phi^{M}(h) + \phi^{DF}(z)]\}\right|^2 \tag{2}$$

The amplitude part A is a constant to maximize light throughput. The phase part of the pupil function consists of two components. $\phi^{M(h)}$ is from the phase mask, which is a function of the mask height map h. $\phi^{DF(z)}$ is from depth defocus, which is a function of the scene depth z. The complete derivation can be found in the supplemental material.

Coded pattern formulation. To simulate the coded pattern in the projector view $I_p^B(h)$, we separate the sharp pattern I_p^S based on the discretized the depth map z_p (21 layers in this paper), convolve with corresponding PSFs, and combine them together. The formula is written as follows.

$$I_p^B(h) = \sum_{z_p} I_p^S(z_p) * PSF(h, z_p) \tag{3}$$

As a consequence, the final image is a differentiable function with respect to the phase height map h, which is the optical parameter that we need to optimize during the training stage.

Geometry dependence. The intensity of the coded pattern is also affected by the scene geometry. Assuming the scene is Lambertian, the reflected intensity depends on the orientation of the surface with respect to the projector θ as well as the distance to the scene $d = \sqrt{x^2 + y^2 + z^2}$. The final intensity should be scaled as follows,

$$I_p^B(x,y) \sim \frac{\cos(\theta(x,y))}{d(x,y)^2} \tag{4}$$

3.3 Image warping

Once we have the image in the projector view I_p^B, we can synthesis the corresponding image in the camera view I_c. This geometry-based image warping has been widely applied for unsupervised depth estimation from stereo pairs [10] and video sequences [45] in a fully differentiable manner.

There are two warping strategies that we can consider:s forward warping $\mathcal{W}^F$ and inverse warping $\mathcal{W}^I$. Forward warping is defined as the mapping from the projector view to the camera view, which requires the depth map in the projector view z_p and the relative pose T_{pc}. Inverse warping is defined as the mapping from the camera view to the projector view, which requires the depth map in the camera view z_c and the relative pose T_{cp}. The intrinsic matrices of the projector and the camera are required for both methods. But these two matrices are fixed and can be calibrated beforehand.

We generate the projector view using inverse warping by adopting the bilinear sampling mechanism proposed in [15]. However, this technique does not correctly

render occluded regions. Thus, we separately generate an occlusion mask using forward warping. Specifically, we warp an all-ones matrix from the projector view to the camera view in the forward mode, and label zero-value pixels as black since there is no light projected on those pixels. The final warping formula is as follows.

$$I_c = \mathcal{W}^I(I_p^B, z_c, T_{cp}) \cdot (\mathcal{W}^F(\mathbf{1}, z_p, T_{pc}) > \epsilon) + \mathcal{U}(0, I_m) + \mathcal{N}(0, \sigma^2) \quad (5)$$

To mimic the noise present in real experiments, we add in uniform distribution between 0 and $I_m = 0.05$ and a Gaussian random variable with $\sigma = 0.005$ to model ambient/global light and read noise, respectively.

3.4 Dataset generation

As discussed in the above sections, to simulate the coded image in camera view I_c accurately, there are three inputs required for a given scene: the depth map from the projector view z_p, the depth map from the camera view z_c, and the relative pose from the projector view to the camera view T_{pc} (T_{cp} is the inverse of T_{pc}). Besides, T_{pc} should be different for different scenes since the camera is freely moving. The most related datasets are for indoor localization or SLAM [24, 26]. However, these datasets are either low resolution, or lack of complex geometries in the foreground, which are not suitable for our task.

Instead, we used the open-source 3D creation suite Blender to generate our own dataset. Different geometric objects with various scaled and orientation are randomly placed in the scene. Given a fixed scene, two depth maps are exported as z_p and z_c, along with the random relative camera pose T_{pc}. The synthetic camera has a 50-mm focal length and a 24mm×36mm sensor. The output depth map is an 800×1200 matrix ranging from 0.7m to 0.95m. And the output camera pose is a 4×4 matrix. The numbers of scenes that we generated for training, validation, and testing elements are 4850, 900, and 200, respectively.

4 Reconstruction algorithm

Given a single captured image in the camera view I_c, the goal is to recover both the 3D point cloud of the scene as well as the camera pose in the projector coordinates. This enables unconstrained and freeform users (cameras) to perform self-localization as well as estimate 3D shape under common coordinates.

The reconstruction pipeline is shown in the right part of Fig. 2. First, we design convolutional neural networks to predict pixel-wise 3D map $(x_p^{cView}, y_p^{cView}, z_p^{cView})$. Although the image is captured from the camera view, the output 3D location should be in the projector coordinate since the pattern is based on the projector. *As a result, the 3D map is with respect to the projector but in the camera view.* Then, we estimate the camera pose using PnP algorithms based on the correspondence between the estimated 3D map and the captured 2D map.

4.1 Image preprocessing

To mitigate the intensity dependency of the surface normal and depth, we apply local normalization (LN) as suggested in [32]:

$$I_c^{LN} = \mathrm{LN}(I_c, x, y) = \frac{I_c(x, y)}{\mu_I(x, y) + \epsilon} \tag{6}$$

$\mu_I(x, y)$ denotes the mean in a small region (17×17 in our simulation) around (x, y), and ϵ is a constant to avoid numerical instabilities.

4.2 Reconstruction network

Empirically, we observed that having one network for x, y estimation (XYNet), and one network for z estimation (ZNet) provide the best performance. The main reason is that x, y localization focuses on global features, while z localization is based on local blur and distortion. ZNet directly outputs the absolute depth values, and XYNet first outputs the relative angles (i.e., x/z and y/z) and then convert to the absolute x, y position by multiplying the ground truth depth. In this way, XYNet only needs to predict relative a 2D position without the dependency on the depth. Both XYNet and ZNet are similar to UNet [33], which is designed as an encoder-decoder architecture with skip connections. The detailed parameters are listed in the supplementary material.

4.3 Loss function

In the input image I_c, there are occluded regions containing no information about the scene. Those regions are masked out from the loss to force the networks to learn only from the patterns.

Our loss function is composed of three individual losses: a root-mean-square (rms) L_{rms} on x, y, z, a gradient loss L_{grad} and a reprojection loss L_{rp}.

$$L = \lambda_1 L_{rms} + \lambda_2 L_{grad}^z + \lambda_3 L_{rp} \tag{7}$$

L_{rms} is a combination of $L_{rms}^x, L_{rms}^y, L_{rms}^z$ to directly force the networks to learn the correct estimation. The gradient loss L_{grad}^z is applied on the depth map to emphasize the network to learn sharp depth boundaries which is common in the natural scene.

$$L_{grad}^z = \frac{1}{\sqrt{N}} \left(\left\| \frac{\partial z_p^{cView}}{\partial x} - \frac{\partial \widehat{z}_p^{cView}}{\partial x} \right\|_2 + \left\| \frac{\partial z_p^{cView}}{\partial y} - \frac{\partial \widehat{z}_p^{cView}}{\partial y} \right\|_2 \right) \tag{8}$$

In our system, the depth information can be extracted from not only the pattern defocus, but also the pattern perspective distortion since the camera and the project are not co-located. To unitize the perspective distortion for depth estimation, we add the reprojection loss L_{rp} between the actual image I_c and

the predicted image $\widehat{I}_c$ from $\widehat{z}_p^{cView}$. The mathematical derivation of $\widehat{I}_c(\widehat{z}_p^{cView})$ can be found in the supplementary material.

$$L_{rp} = \frac{1}{N} \left\| I_c - \widehat{I}_c(\widehat{z}_p^{cView}) \right\|_1 \tag{9}$$

Here, ℓ_1-norm is used since I_c is sparse.

4.4 Training details

During the training, the input image patch has a size 256×256 px, which is randomly cropped from our dataset mentioned in Sec. 3.4. At test time, since our networks are fully-convolutional, images size can be any multiple of 16. We train the parameters of the optical system (i.e., the mask height map) jointly with the digital convolutional layers. Empirically, we find that the result converges better by training in two stages. First, we pre-train the mask height map and ZNet with L^z_{rms} and L_{grad} in a colocated setting where T_{pc} is identity. Second, we train the entire model using all losses end-to-end.

4.5 Camera pose estimation

Our networks output the 3D coordinates of the scene from the camera's point of view. We can then calculate the camera pose by passing the 3D coordinates and the corresponding 2D local image coordinates to a PnP solver. We use OpenCV [2] implementation of PnP solver [8] made robust with RANSAC [39].

Conceptually, the (x, y) locations provided XYNet relies on analyzing the spatial distribution of the Kronecker multiplexed pattern. This means that a sufficiently large receptive field is required to estimate (x, y) accurately. However, in regions with small features and significant depth variations, the projected pattern is highly distorted, yielding erroneous (x, y) estimates. Assuming that the majority of the scene is smooth without rapid depth variations, a robust PnP solver can estimate the camera pose accurately.

Refinement of (x, y). While the estimation of (x, y) might not be good for specific regions with small features and large depth variations, the z estimation is less affected since ZNet extracts local blurring information. Thus, (x, y) is further refined using the z estimation and the robustly estimated camera pose.

5 Simulation results

5.1 Optimized mask design and testing results

The top left of Fig. 3 shows the optimized phase mask height map that we obtain from the training procedure. The corresponding PSFs at different depth ranges are shown below. At the focused depth (0.8m), the PSF is a dot. As the depth reduces, it splits to two dots vertically. As the depth increases, it splits to two dots horizontally. This variation makes the robust depth estimation possible.

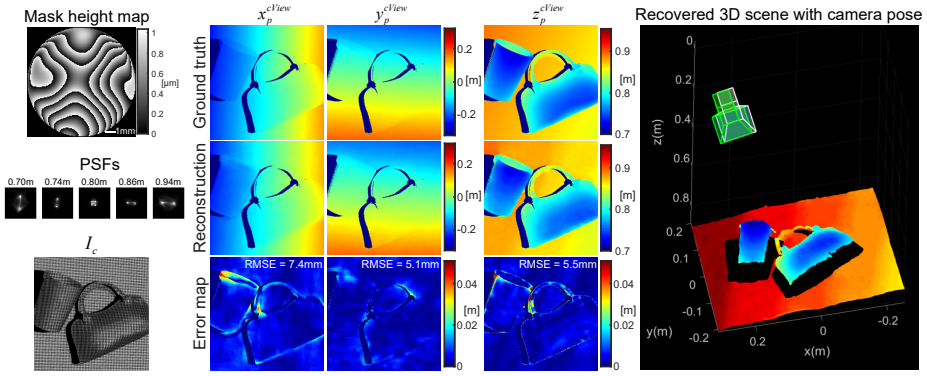

Fig. 3. Simulation results. (Left) The learned phase mask and its corresponding PSFs at different depths. I_c is an example of the input image in simulation. (Center) The output of XYNet and ZNet, containing the 3D map in the projector coordinate. (Right) The estimated point cloud of the scene in the projector coordinate. The estimated camera pose (white) is close to the ground truth (green).

To evaluate the performance in simulation, we show the reconstruction results of a testing scene - a cup and a handbag on a tilted floor (Fig. 3). The camera captures the scene with the projector pattern as I_c. The trained networks output the 3D location in the project coordinate for each pixel. Comparing with the ground truth, the error is mainly near the depth boundary. The 3D point cloud is shown in the right part of Fig.. 3. The estimated camera pose (color in white) is also shown in the figure, which is close to the ground truth (color in green). The error in translation is $(0.013, 0.009, 0.016)$ meters, and the error in rotation is $(0.013, 0.013, 0.002)$ radians for pitch, yaw and roll. This example demonstrates that we are able to accurately output the 3D scene as well as the camera pose from just a single shot. More analysis can be found in the supplementary materials.

5.2 Ablation study and comparisons

There are two important components in our projector system, the projector pattern and the inserted phase mask. The results of the ablation study are shown in Table 1. Although there are various single-shot patterns, many are not suitable for comparison because our system requires the pattern to contain global context with dense local features. For example, test A shows that a uniform grid pattern is not able to provide the spatial uniqueness to give the x and y locations. And the results from Kinect [28] and M-array [20] patterns are still worse than our proposed Kronecker-multiplexed pattern. On the other hand, test D shows that the depth estimation error increases dramatically when there is no mask. In this case, the PSF becomes a disk function, and is identical at both sides of the focal plane, which is hard to estimate depth from the pattern.

Test	Projector pattern	Phase mask	L^x	L^y	L^z
A	Grid	Optimized	52.1	50.6	8.7
B	Kinect	Optimized	15.8	18.1	9.2
C	M-array	Optimized	12.9	15.4	15.5
D	Kronecker-multiplexed	No mask	8.8	10.3	90.3
E	**Kronecker-multiplexed**	**Optimized**	**8.3**	**10.1**	**6.1**

Table 1. Ablation study (the unit of all the losses is mm)

Model	Projector pattern	L_c^z	L_c^z with camera misalignment
A	FreeCam3D	7.6	7.8
B	Kinect + UNet	7.8	15.7
C	Kinect + DispNet	7.4	14.8

Table 2. Model comparison (the unit of all the losses is mm)

We further compare our method with recent deep learning-based algorithms for SL system. Since these algorithms require the camera to be pre-calibrated and fixed, we generate another dataset with a fixed camera pose (10 cm baseline). Model B is trained with UNet [33] and Kinect pattern, and model C is trained with DispNet [25, 32] and Kinect pattern. L_c^z is the rms loss on recovered depth in camera coordinate. As shown in the Table 2, our system has a similar performance when the camera is well-calibrated and is more robust when the camera pose is misaligned (12cm baseline).

6 Experiment results

Experimental setup. A picture of the setup is shown in the left part of Fig. 4. We use an Epson VS355 LCD Projector (1280×800, 10μm pixel size) with a 50-mm $f/1.8$ standard prime lens. The phase mask is fabricated by the Reactive Ion Etching (RIE) process. The diameter of the mask is 10.5mm with 70-μm pixel size. The projector only projects green patterns, which mitigates the PSFs' dependency on wavelength. The projector PSFs are calibrated experimentally for any fabrication imperfection and system misalignment. The calibration process can be found in the supplementary material. The networks are fine-tuned based on the experimental PSFs.

At the camera side, our sensor is a 5472×3648 machine vision color camera (BFSPGE-200S6C-C) with 2.4μm pixel size. To match the pixel size of the projector, the captured image is rescaled to the resolution of 1312×864. The imaging lens is a 50-mm $f/16$ lens. The use of a small aperture in the camera makes its depth field very large, and hence its PSF is near-invariant in our operating depth range.

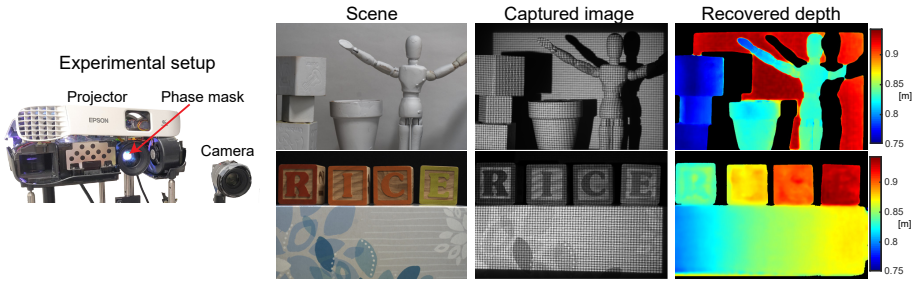

Fig. 4. Experimental setup and results for static scenes. (upper row) complicated scene and (bottom row) texture scene.

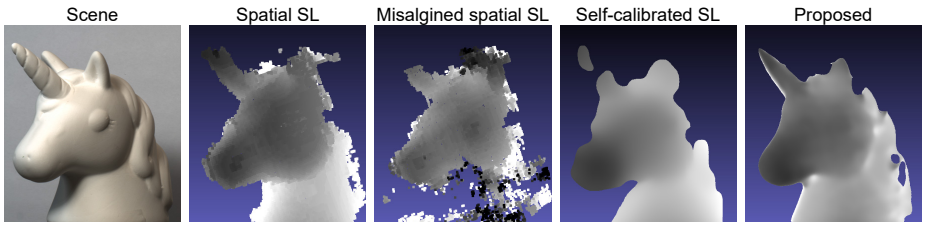

Fig. 5. Experimental depthmap comparisons with single-shot structured light methods.

As we report in the supplementary material, the rms error in depth estimation is 3.7 mm for 0.7m - 0.95m range.

Static scenes. We demonstrate the results for static scenes with a fixed camera pose. Fig. 4 shows the recovered depth maps $\hat{z}_p^{cView}$. Our algorithm recovers depth for both textureless scene and textured scene (with finetuning using the same dataset with random texture). By combining with the estimated $(\hat{x}_p^{cView}, \hat{y}_p^{cView})$, we show an example of the recovered 3D point cloud and camera pose in Fig. 1.

Comparisons with related SL systems. To confirm the effectiveness of our method, we compared our technique to related single-shot SL methods (Fig. 5). The baseline for all the methods is 10 cm. For spatial-coding SL, we use a pseudo-random dot pattern with the Kinect v1 stereo matching algorithm [28]. To further test the sensitivity of this method to calibration, we recover the shape after adding a slight error in the rotation angle between the projector and the camera (0.2 degrees). As we can see, even a small misalignment affects the result significantly. On the other hand, there are self-calibrating single-shot scanning techniques. Here we implemented one using markers [7]. Although the 3D shape was recovered, the resolution is extremely low. This is because the pattern for self-calibrating SL is sparse in order to find correspondences in a practical manner without the help of the epipolar constraint. Since only low

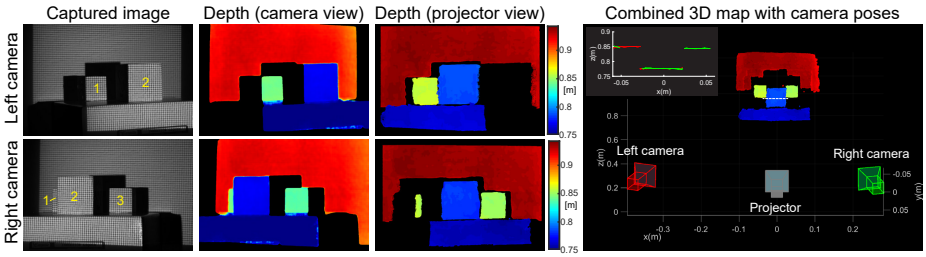

Fig. 6. 3D reconstruction from two cameras. Each camera only sees a part of the scene. Since our system estimates the 3D map in world coordinates, those two point clouds can be combined seamlessly. The height along the dashed scanline is plotted.

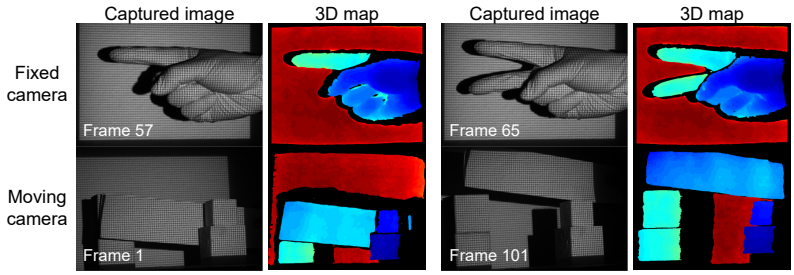

Fig. 7. 3D reconstruction with dynamic scenes.

resolution is recovered, all the high-frequency shapes such as the horn of the unicorn object cannot be recovered and surfaces are all smoothed out.

Overall, our proposed method provides comparable depth resolution with the spatial SL and better shape boundaries. While the spatial SL is sensitive to the calibration misalignment, our technique does not require calibration, which is one important strength of our algorithm.

Multi-camera systems. One advantage of our method is that the output 3D point cloud is in world coordinates. If multiple cameras are looking at the same scene from different perspectives, their results can be directly combined to create a complete reconstruction of the scene. Fig. 6 gives a sample scene with three cubes on a table. Each camera only a part of the scene. However, we can observe the similarity in the 3D map in regions that are in both views. All three cubes are visible in the combined 3D map as shown on the right side of the figure. As a byproduct, the left and right camera poses are estimated. This example demonstrates interesting applications that involve multiple participants in a shared scene.

Dynamic scene and moving camera. Since our method is single-shot, it offers the ability to work for dynamic scenes with moving cameras. In Fig. 7,

the first row shows hand gestures captured from a fixed camera. The second row shows a paper stripe is swiped from the bottom to the top, while the camera is shifted to the left. The full 3D reconstruction with camera pose can be found in supplementary materials. All videos are recorded in a 30Hz frame rate.

7 Discussion and Conclusion

In this paper, we demonstrate FreeCam3D for 3D imaging of the scene where the camera is not constrained to a rigid position relative to the projector. From a single image captured by the camera, we estimate the 3D map of the scene and the camera pose. These coordinates are in the shared world coordinates, represented by the fixed projector. We built a prototype and demonstrated high-quality 3D reconstruction with an unconstrained camera.

Practically, we envision that our system will be implemented using NIR lighting, like the Microsoft Kinect, so as to not interfer with human vision. Most visible-light textures, which are from dyes, are practically transparent in NIR, and are low-contrast. In such case, the texture on scene will be dominated by the structured light. Therefore, we focus most of our results on texture-less scenes. Finally, texture dependency can be mitigated by enhancing the training pipeline to include textures. We provide such analysis in the supplementary materials.

Limitations. The advances made by the proposed system come with certain limitations. First, the 3D estimates of our technique are of lower spatial resolution than what can be obtained with traditional structured light systems with a similar camera and projector; this can be attributed to many sources including the use of defocus blur for depth, the loss in resolution due to the design of the projected pattern and, finally, the lack of knowledge of the pose of the camera. Second, the intended applications of our system are in enabling shared spaces that facilitate interaction of multiple participants in an AR/VR setting. To ultimately realize such an environment, we also need to increase the field-of-view (FoV) of the system. Our work can be extended to an increased FoV by installing multiple fixed projectors, each with its optimized phase mask. The projectors can be pre-calibrated with respect to each other, while the participants with cameras can move around in an unconstrained fashion. Regions of occlusions can also be dealt with a multi-projector system. Finally, our experimental results are captured in the visible light with texture-less objects. When using this technique in a real application, the system can be implemented using near-infrared (NIR) light and reap the dual benefits of being non-intrusive to human vision and making most objects texture-less.

Acknowledgement

This work was supported in part by NSF grants IIS-1652633 and CCF-1652569, DARPA NESD program N66001-17-C-4012, and JSPS KAKENHI grants JP20H00611 and JP16KK0151.

References

1. Benveniste, R., Ünsalan, C.: A color invariant based binary coded structured light range scanner for shiny objects. In: International Conference on Pattern Recognition (ICPR). pp. 798–801 (2010)
2. Bradski, G., Kaehler, A.: Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc." (2008)
3. Chakrabarti, A.: Learning sensor multiplexing design through back-propagation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 3081–3089 (2016)
4. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 10193–10202 (2019)
5. Farid, H., Simoncelli, E.P.: Range estimation by optical differentiation. Journal of the Optical Society of America A (JOSA A) **15**(7), 1777–1786 (1998)
6. Furukawa, R., Nagamatsu, G., Kawasaki, H.: Simultaneous shape registration and active stereo shape reconstruction using modified bundle adjustment. In: International Conference on 3D Vision (3DV). pp. 453–462 (2019)
7. Furukawa, R., Naito, M., Miyazaki, D., Baba, M., Hiura, S., Sanomura, Y., Tanaka, S., Kawasaki, H.: 3d endoscope system using asynchronously blinking grid pattern projection for hdr image synthesis. In: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE), pp. 16–28 (2017)
8. Gao, X.S., Hou, X.R., Tang, J., Cheng, H.F.: Complete solution classification for the perspective-three-point problem. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **25**(8), 930–943 (2003)
9. Girod, B., Scherock, S.: Depth from defocus of structured light. In: Optics, Illumination, and Image Sensing for Machine Vision IV. vol. 1194, pp. 209–215 (1990)
10. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 270–279 (2017)
11. Goodman, J.W.: Introduction to Fourier optics. Roberts and Company Publishers (2005)
12. Guo, Q., Alexander, E., Zickler, T.: Focal track: Depth and accommodation with oscillating lens deformation. In: IEEE International Conference on Computer Vision (ICCV). pp. 966–974 (2017)
13. Haim, H., Elmalem, S., Giryes, R., Bronstein, A.M., Marom, E.: Depth estimation from a single image using deep learned phase coded mask. IEEE Transactions on Computational Imaging (TCI) **4**(3), 298–310 (2018)
14. Hitoshi, M., Hiroshi, K., Ryo, F.: Depth from projector's defocus based on multiple focus pattern projection. IPSJ Transactions on Computer Vision and Applications (CVA) **6**, 88–92 (2014)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 2017–2025 (2015)
16. Kawasaki, H., Furukawa, R., Sagawa, R., Yagi, Y.: Dynamic scene shape reconstruction using a single structured light pattern. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. Ieee (2008)
17. Kawasaki, H., Horita, Y., Masuyama, H., Ono, S., Kimura, M., Takane, Y.: Optimized aperture for estimating depth from projector's defocus. In: International Conference on 3D Vision (3DV). pp. 135–142 (2013)

18. Kawasaki, H., Horita, Y., Morinaga, H., Matugano, Y., Ono, S., Kimura, M., Takane, Y.: Structured light with coded aperture for wide range 3D measurement. In: IEEE Conference on Image Processing (ICIP). pp. 2777–2780 (2012)
19. Lee, J., Gupta, M.: Stochastic exposure coding for handling multi-tof-camera interference. In: IEEE International Conference on Computer Vision (ICCV). pp. 7880–7888 (2019)
20. Lei, Y., Bengtson, K.R., Li, L., Allebach, J.P.: Design and decoding of an m-array pattern for low-cost structured light 3d reconstruction systems. In: IEEE International Conference on Image Processing (ICIP). pp. 2168–2172 (2013)
21. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o (n) solution to the pnp problem. International Journal of Computer Vision (IJCV) **81**(2), 155 (2009)
22. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Transactions on Graphics (TOG) **26**(3), 70 (2007)
23. Li, Q., Biswas, M., Pickering, M.R., Frater, M.R.: Accurate depth estimation using structured light and passive stereo disparity estimation. In: IEEE International Conference on Image Processing (ICIP). pp. 969–972 (2011)
24. Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., Leutenegger, S.: Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. arXiv:1809.00716 (2018)
25. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048 (2016)
26. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. arXiv:1612.05079 (2016)
27. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep optics for single-shot high-dynamic-range imaging. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1375–1385 (2020)
28. Microsoft: Xbox 360 Kinect (2010), http://www.xbox.com/en-US/kinect
29. Microsoft: Kinect for Windows (2013), http://www.microsoft.com/en-us/
30. Nayar, S., Watanabe, M., Noguchi, M.: Real-time focus range sensor. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **18**(12), 1186–1198 (1996)
31. Pavani, S.R.P., Thompson, M.A., Biteen, J.S., Lord, S.J., Liu, N., Twieg, R.J., Piestun, R., Moerner, W.: Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. Proceedings of the National Academy of Sciences (PNAS) **106**(9), 2995–2999 (2009)
32. Riegler, G., Liao, Y., Donne, S., Koltun, V., Geiger, A.: Connecting the dots: Learning representations for active monocular depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7624–7633 (2019)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI). pp. 234–241 (2015)
34. Salvi, J., Fernandez, S., Pribanic, T., Llado, X.: A state of the art in structured light patterns for surface profilometry. Pattern Recognition **43**(8), 2666–2680 (2010)
35. Shechtman, Y., Sahl, S.J., Backer, A.S., Moerner, W.: Optimal point spread function design for 3d imaging. Physical Review Letters (PRL) **113**(13), 133902 (2014)

36. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM Transactions on Graphics (TOG) **37**(4), 1–13 (2018)
37. Sun, Q., Tseng, E., Fu, Q., Heidrich, W., Heide, F.: Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1386–1396 (2020)
38. Tang, S., Zhang, X., Tu, D.: Fuzzy decoding in color-coded structured light. Optical Engineering **53**(10), 104104 (2014)
39. Torr, P.H., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding (CVIU) **78**(1), 138–156 (2000)
40. Ulusoy, A.O., Calakli, F., Taubin, G.: Robust one-shot 3d scanning using loopy belief propagation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 15–22 (2010)
41. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Transactions on Graphics (TOG) **26**(3), 69 (2007)
42. Watanabe, M., Nayar, S.K.: Rational filters for passive depth from defocus. International Journal of Computer Vision (IJCV) **27**(3), 203–225 (1998)
43. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: Phasecam3d—learning phase masks for passive single view depth estimation. In: IEEE International Conference on Computational Photography (ICCP). pp. 1–12 (2019)
44. Zhang, X., Li, Y., Zhu, L.: Color code identification in coded structured light. Applied Optics **51**(22), 5340–5356 (2012)
45. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1851–1858 (2017)