Single-shot Monocular RGB-D Imaging using Uneven Double Refraction

Andreas Meuleman[1]* Seung-Hwan Baek[1]*† Felix Heide[2] Min H. Kim[1]

[1]KAIST [2]Princeton University

Abstract

Cameras that capture color and depth information have become an essential imaging modality for applications in robotics, autonomous driving, virtual, and augmented reality. Existing RGB-D cameras rely on multiple sensors or active illumination with specialized sensors. In this work, we propose a method for monocular single-shot *RGB-D imaging. Instead of learning depth from single-image depth cues, we revisit double-refraction imaging using a birefractive medium, measuring depth as the displacement of differently refracted images superimposed in a single capture. However, existing double-refraction methods are orders of magnitudes too slow to be used in real-time applications, e.g., in robotics, and provide only inaccurate depth due to correspondence ambiguity in double reflection. We resolve this ambiguity optically by leveraging the orthogonality of the two linearly polarized rays in double refraction – introducing* uneven double refraction *by adding a linear polarizer to the birefractive medium. Doing so makes it possible to develop a real-time method for reconstructing sparse depth and color simultaneously in real-time. We validate the proposed method, both synthetically and experimentally, and demonstrate 3D object detection and photographic applications.*

1. Introduction

RGB-D cameras that simultaneously acquire color and depth information have emerged as a critical imaging modality for applications in computer vision and graphics, including autonomous driving, robotics, photography, and mixed reality. However, broadly adopted RGB-D cameras either rely on multiple cameras [18] or combine a conventional camera with a separate depth sensor. These latter typically rely on an active illumination module that modulates light either spatially [23, 14, 11] or temporally [15], such as a time-of-flight (TOF) camera. Existing approaches to monocular RGB-D imaging, i.e., using only a single camera, aim to recover depth-from-defocus [28, 17], depth-from-focus [5], and depth-from-refraction [20, 4, 7, 1]. Although all of these methods rely only on a conventional camera with small footprint and cost, they require *multiple shots* to obtain depth, which prohibits their use in dynamic real-world scenes.

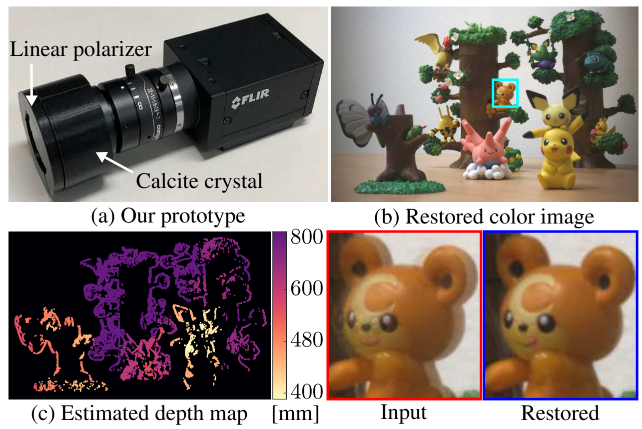


Figure 1. (a) Our prototype, consisting of a linear polarizer and a calcite crystal in front of a conventional camera. (b) and (c) estimated depth map and restored color image from the input. Our algorithm provides both sparse depth and clear RGB image within 34 ms. Refer to the supplemental material for real-time demo.

The ultimate goal of monocular RGB-D imaging is to obtain color and depth simultaneously from a single shot. Plenoptic imaging, i.e., light-field imaging, approaches this problem by combining the objective lens with a micro-lens array in front of the sensor to capture multi-perspective sub-images of a scene. Unfortunately, this angular resolution comes at the cost of a loss in spatial resolution, and the depth range is fundamentally limited by the short baseline [19]. Alternative approaches relying on pixel arrays that alternately see half of the aperture [30, 8], thereby capturing subsampled stereo views, suffer from a narrow baseline at the long distances that TOF RGB-D cameras excel at.

To tackle all of the above limitations, instead of separating angular measurements, we superpose them by revisiting depth-from-double-refraction, while lifting existing ambiguity and runtime restrictions of double-refraction methods. Baek et al. [1] use a birefringent medium in front of the camera lens, such as a calcite crystal, to overlap two shifted images that encode depth via their local disparity. However, the intensities of these two images are identical. This fundamental ambiguity of searching stereo correspondences in double refraction images results in very low computational

∗: Equal contribution. †: Now at Princeton University efficiency, with more than half a minute compute time for single RGB-D frame, and inaccurate depth estimates, prohibiting real-time RGB-D imaging applications.

In this work, we introduce a real-time single-shot monocular RGB-D camera. Specifically, we make the following contributions:

To tackle double-refraction ambiguity, we exploit the optical phenomenon that each refraction is linearly polarized in double refraction by a birefringent medium and that these two refractions are orthogonal to each other. This allows us to optically control to make the ratio of each polarimetric refraction uneven, by combining a linear polarizer with the birefringent medium. This *uneven double refraction* resolves the ambiguity of correspondence in depth-from-double-refraction. We present a novel *joint reconstruction method for depth and color*. Our key idea is to restore a clear color image using only the higher intensity by iteratively eliminating the refraction of the weaker intensity in uneven double refraction, while estimating depth from displacement in double refraction. Building on this resolved ambiguity in the image formation, we achieve real-time RGB-D acquisition by devising a novel *rectification method for double-refraction images* achieving a speedup of over a factor of 1000 over state-of-the-art methods. This feat allows us to acquire high-quality depth and color with real-time performance on consumer GPU hardware. Figure 1 shows our prototype and a captured RGB-D image.

We validate our method synthetically and on experimental data, where our approach outperforms state-of-the-art monocular RGB-D methods in accuracy, depth-range, and runtime. We demonstrate a variety of applications using the proposed RGB-D imager, including 3D object detection and photographic applications. All codes, models, and detailed optical designs are published to ensure reproducibility (`https://github.com/KAIST-VCLAB/fastbirefstereo.git`).

2. Related Work

In this section, we discuss existing single-shot monocular RGB-D imaging methods.

Depth from Defocus Depth information can be estimated by analyzing the level of defocus in the image [24], i.e., the distance is proportional to the amount of blurriness. However, due to the low-frequency nature of defocus blur, its depth cues often are not sufficient to provide accurate depth. Changing the shape of the aperture [17, 2, 34] and employing a mask on the sensor [29] improves depth estimation over isotropic kernels; however, such approaches still provide inaccurate depth and color due to the fundamentally low-frequency depth cues. The proposed method utilizes uneven double refraction as a high-frequency depth cue, allowing for improved depth and high-quality color images.

Depth from Light Field Light fields contain subimages with short baselines that allow for depth estimation. Existing methods make use of disparity among subimages in horizontal and vertical directions to estimate depth [19]. Wang et al. [31, 32] account for occlusion to estimate sharp depth transition around edges. However, existing light field cameras need to be equipped with a lenticular lens array, fundamentally limiting the spatial resolution as a tradeoff for angular resolution.

Recently, a reduction of this concept to subsampled stereo images has been proposed to estimate depth, using customized dual pixels sensors [30, 8]. In this approach, the micro-lens array on the sensor is used similarly to the lenticular lens in a light field camera. Specifically, pixels alternately block half of the aperture by blocking light in half of a pixel's active area, resulting in subsampled stereo views. However, the disparity range of this dual-pixel sensing is limited to a few pixels. In contrast, the proposed method uses an unmodified conventional sensor, and our birefringent medium provides large disparity ranges of more than 20 pixels allowing for larger depth ranges.

Depth from Reflection Double reflection methods capture depth using a slanted mirror in front of the camera [26, 33]. This approach requires a very large mirror, sacrificing mobility due to the large form factor. Different from depth from reflection, our imaging setup consists of only one camera with two flat optical materials, a birefringent medium, and a linear polarizer, making the system compact and maintaining the optical axis of the original camera.

Depth from Refraction Traditional depth-from-refraction methods [20, 4, 6, 7] estimate depth from the displacement of multiple differently refracted images. Besides, specialized imaging setups with an optical component, such as a prism or a micro-lens array, have been devised to capture depth from a single-shot input. Lee et al. [16] installs two prisms of a camera to capture two perspective images in a single shot, at the cost of sacrificing sensor resolution and high-quality imagery. Baek et al. [1] propose to estimate depth from double refraction. However, due to the intrinsic ambiguity of two displaced images with equal intensities, the accuracy of the reconstructed depth and color image is fundamentally limited, and complex recovery methods require more than half a minute per single image. In contrast, we rely on the cross linear polarization states of the displaced images and attenuate one displaced component by an additional polarizer, resolving the ambiguity and enabling efficient recovery of both depth and color.

Learning Depth from a Single Image Many recent works have explored learning depth from a single image depth cues, such as defocus, perspective, and parallax, using neural networks [24, 9, 13]. While demonstrating remarkable results, such approaches still suffer low accuracy and do not generalize across cameras and scene semantics, e.g., outdoor versus indoor. In contrast, our method uses optically encoded disparity from uneven double refraction to *measure* depth, instead of learning it from indirect depth cues.

3. Uneven Double Refraction

Optical Design In double refraction, a pair of rays, the corresponding ordinary ray (o-ray) and extraordinary ray (e-ray) generate shifted copies of the same latent scene image, which are captured as superposition. These rays have equal intensities for typical unpolarized natural incident light, creating *ambiguity* in determining whether an edge is generated by the o-ray or the e-ray. While existing depth-from-double-refraction methods [1] partially address this issue with the sophisticated optimization methods that use the dual cost function in the image gradient domain, such computationally expensive algorithms prohibit real-time processing and are fundamentally limited in depth and image quality by the double refraction ambiguity.

In contrast, we propose to *optically* resolve this ambiguity by exploiting the fact that the o-ray and the e-ray are linearly polarized and perpendicular to each other. Owing to the polarimetric properties of o-/e-ray, we can control the intensity proportion of both light rays by combining a linear polarizer and a birefringent medium, see Figure 2. Specifically, we adjust the angle of polarizer so that the e-ray becomes attenuated with the lower intensity and can be effectively removed with the proposed reconstruction method.

Image Formation Next, we describe the image formation model for *uneven double refraction*. Assuming a pinhole camera model with focal length f, a light ray from scene point P_s is projected to the direct pixel P_d if there is no birefringent medium, as shown in Figure 2. Once a birefractive material that exhibits optical anisotropy to the polarization states of light waves, e.g., a calcite crystal, is placed in the light path and an unpolarized incident ray passes through the medium, this ray is split into two, which have different directions of propagation, with the o-ray following Snell's law and the e-ray violating Snell's law. The o-ray and the e-ray follow different paths and project to pixels P_o and P_e, respectively. To achieve uneven double refraction, we place a linear polarizer in front of the medium. The rotation of this polarizer adjusts the ratio between the o-ray and the e-ray. Note that we assume the linear polarizer is thin enough not to refract the light rays. Birefractive disparity $r_{\overline{oe}}$ is then defined as the displacement vector from P_o to P_e. Figure 2 shows the optical light transport of polarized double refraction in our setup.

Although birefractive disparity exists in both horizontal and vertical directions [1], we assume a rectified birefractive disparity image in the following. To this end, we introduce an efficient novel rectification method later in this

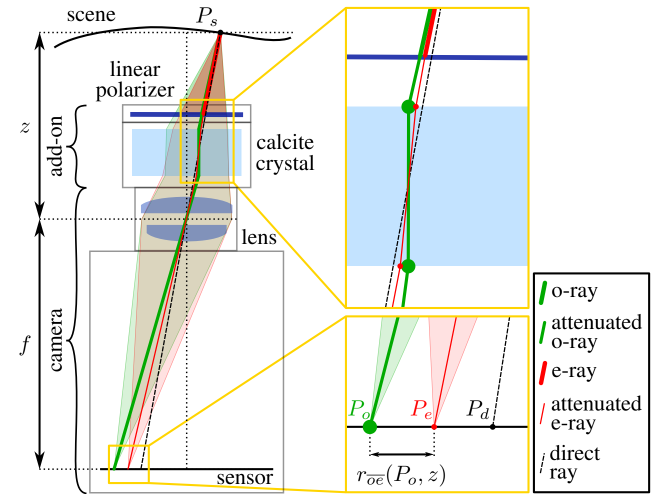


Figure 2. Our image formation for uneven double refraction. A scene point P_s is directly projected to P_d, while o-ray and e-ray of P_s are captured at different points P_o and P_e, respectively. Using the linear polarizer, e-ray has less intensity than o-ray. We estimate depth from birefractive disparity $r_{\overline{oe}}$.

work. The rectified image formed by o-ray and e-ray, $\tilde{I}_o$ and $\tilde{I}_e$, are displaced with the rectified birefractive disparity $\tilde{r}_{\overline{oe}}(z)$. The intensities of $\tilde{I}_o$ and $\tilde{I}_e$ are also different because of the linear polarizer, which introduces uneven double refraction with an amount of τ, obtained through calibration (refer to the supplemental document for more details). Therefore, the e-ray image $\tilde{I}_e$ can be formulated as: $\tilde{I}_e = \tau A(\tilde{I}_o, \tilde{r}_{\overline{oe}}(z))$, where $A(\tilde{I}_o, \tilde{r}_{\overline{oe}}(z))$ is a function that translates the o-ray image $\tilde{I}_o$ according to the disparity $\tilde{r}_{\overline{oe}}(z)$. The captured image $\tilde{I}_c = \tilde{I}_o + \tilde{I}_e$ can be reformulated as a superimposition of the o-ray image and the transformed o-ray image, corresponding to the e-ray image:

$$\tilde{I}_c = \tilde{I}_o + \tau A\left(\tilde{I}_o, \tilde{r}_{\overline{oe}}(z)\right). \tag{1}$$

4. Joint Reconstruction of Color and Depth

Given an uneven rectified input image, we propose an efficient and effective joint depth and color reconstruction method. We devise a non-blind color restoration method that can efficiently and effectively remove uneven double refraction. The key idea here is to iteratively eliminate the weak refraction component (the e-ray image) from the uneven double refraction. Analogous to the concept of the cost volume in traditional stereo imaging [12], we use our non-blind color restoration method to calculate a restoration volume that stores a set of restored color images for every depth candidate. Then, we estimate the sparse depth by selecting among the restoration candidates the depth at which color reconstruction is optimal. Note that obtaining a clean color image is the byproduct of this depth estimation.

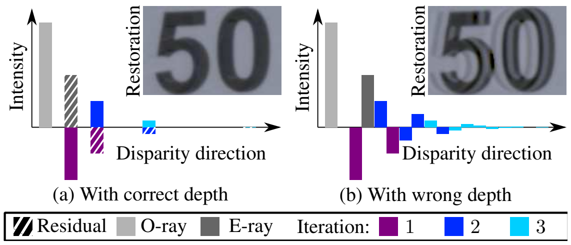


Figure 3. (a) We remove the e-ray component of the gray slashed bar by using the two purple bars generated by translation by known $\tilde{r}_{\overline{oe}}(z)$ and scaled by τ. This causes a new residual error denoted as the purple slashed bar, which can also be removed similarly. We repeat this process until the intensity of the residual error becomes lower than a threshold. (b) Given a wrong disparity $\tilde{r}_{\overline{oe}}(z)$, we cannot correctly remove residuals, yielding restoration artifacts.

4.1. Color Image Restoration

The goal of our color image restoration is to recover the latent o-ray image $\tilde{I}_o$ from the captured image $\tilde{I}_c$. Since depth is also as yet unknown, we restore images for all depth candidates z within a range resulting in a *restoration volume* that contains image restoration values $\hat{I}_o^z$ for each depth candidate. Next, we describe our image restoration method for depth candidate z.

Using our birefractive model, we first compute the corresponding birefractive disparity for z (Equation (9)). The key idea of our image restoration is to iteratively remove the e-ray intensity, which is weaker than that of o-ray, from the captured image $\tilde{I}_c$.

We denote the current restoration of $\tilde{I}_o$ at the n-th iteration as $\hat{I}_o^{z,(n)}$. For initialization at the first iteration, we start with the captured input image: $\hat{I}_o^{z,(0)} = \tilde{I}_c$. Next, we define a residual image that we want to remove from the current estimate: $\Delta^{z,(0)} = \hat{I}_o^{z,(0)} - \tilde{I}_o = \tau A(\tilde{I}_o, \tilde{r}_{\overline{oe}}(z))$. However, the residual $\Delta^{z,(0)}$ cannot be calculated directly because the ground truth $\tilde{I}_o$ is also unknown. We therefore compute an approximated residual image $\hat{\Delta}^{z,(0)}$ using our current estimate $\hat{I}_o^{z,(0)}$ instead of the ground truth $\tilde{I}_o$; this is similar to the mirror-reflection calculation by Yano et al. [33]: $\hat{\Delta}^{z,(0)} = \tau A(\hat{I}_o^{z,(0)}, \tilde{r}_{\overline{oe}}(z)) = \tau A(\tilde{I}_o, \tilde{r}_{\overline{oe}}(z)) + \tau^2 A(\tilde{I}_o, 2\tilde{r}_{\overline{oe}}(z))$. We then update the current estimate of the o-ray image by subtracting the approximated residual:

$$\hat{I}_o^{z,(1)} = \hat{I}_o^{z,(0)} - \hat{\Delta}^{z,0} = \tilde{I}_o - \tau^2 A\left(\tilde{I}_o, 2\tilde{r}_{\overline{oe}}(z)\right). \quad (2)$$

As the attenuation ratio of e-ray τ is by definition less than one, the new residual $\Delta^{z,(1)} = -\tau^2 A(\tilde{I}_o, 2\tilde{r}_{\overline{oe}}(z))$ has a lower intensity level than that of the previous residual $\Delta^{z,(0)}$, making our current estimate $\hat{I}_o^{z,(1)}$ closer to the ground truth than the previous estimate $\hat{I}_o^{z,(0)}$. In the next iteration, the approximated residual is similarly defined as follows: $\hat{\Delta}^{z,(1)} = -\tau^2 A(\hat{I}_o^{z,(1)}, 2\tilde{r}_{\overline{oe}}(z)) = -\tau^2 A(\tilde{I}_o, 2\tilde{r}_{\overline{oe}}(z)) + \tau^4 A(\tilde{I}_o, 4\tilde{r}_{\overline{oe}}(z))$. The current image estimate is then also updated as: $\hat{I}_o^{(2)} = \hat{I}_o^{z,(1)} - \hat{\Delta}^{z,(1)} = \tilde{I}_o + \tau^4 A\left(\tilde{I}_o, 4\tilde{r}_{\overline{oe}}(z)\right)$. We repeat this process until the intensity level of the approximated residual is less than the threshold. We found that three iterations are sufficient to allow the joint estimation to converge (see Figure 3).

We can see from Equation (2) that the residual error of our algorithm after N iterations is

$$\Delta^{z,(N)} = -\tau^{2^N} \cdot A(\tilde{I}_o, 2^N \cdot \tilde{r}_{\overline{oe}}(z)). \quad (3)$$

Note that our restoration method converges with the speed of τ powered by 2^N, faster than the Taylor expansion [33], which converges at $(-\tau)^{N+1} \cdot A(\tilde{I}_o, (N+1) \cdot \tilde{r}_{\overline{oe}}(z))$ with the same N number of iterations and speed of τ powered by $(N+1)$.

4.2. Depth Estimation

To use double refraction to estimate depth, the existing birefractive stereo method estimates the correspondence between the o-ray and e-ray pixels by defining the cost volume $C^z(P)$ as the similarity of the gradient profiles of the o-ray and e-ray pixels [1]. They calculate the cost volume twice due to the ambiguity in double refraction, and then apply a non-local cost aggregation [36] that also costs as much as the dual cost calculation. This results in high computational cost and not easily parallelizable.

In contrast, by making use of uneven double refraction and the efficient image restoration method, we can estimate depth $Z(P)$ for each pixel P from the restoration volume $\hat{I}_o^z(P)$ by defining a depth cost volume $C^z(P)$ to indicate the cost of selecting depth candidate z for pixel P.

The key insight of our method is that our image reconstruction produces a clear natural image only if the given depth candidate z is correct. Otherwise, the restored image contains multi-refraction artifacts, as shown in Figure 3. This is because wrong depth values cannot correctly remove the image residuals, but instead introduce false edges as artifacts. Therefore, we define the depth cost $C^z(P)$ as the sum of the *gradient magnitudes* of neighboring pixels about P in the restoration volume $\hat{I}_o^z(P)$. The depth cost $C^z(P)$ is defined as follows:

$$C^z(P) = \sum_{P' \in K(P)} \left| \frac{\partial \hat{I}_o^z}{\partial x}(P') \right|, \quad (4)$$

where $K(P)$ is the set of pixels in a window centered at P of size 61×61. We implement this calculation using two linear filters: an efficient Sobel filter for the gradient and the box filter for the neighborhood. Once we have computed the depth cost volume for every depth candidate z, we assign the depth of P so as to minimize the cost:

$$Z(P) = \arg\min_z C^z(P). \quad (5)$$

With the estimated depth Z, we can reconstruct the final color image $\hat{I}_o^Z$ from the restoration volume $\hat{I}_o^z$ (Section 4.1). Note that our estimated depth values are valid around edges, where uneven double refraction is clearly visible and without ambiguity. Therefore, we compute a validity mask so that we can retain only pixels having strong horizontal gradients ($\left|\frac{\partial \hat{I}_o^Z}{\partial x}(P)\right| > \text{Thres}_{\text{grad}}$) and top score among depth candidate in terms of cost ($\max_z C^z(P) - \min_z C^z(P) > \text{Thres}_{\text{cost}}$).

5. Rectification for Double Refraction

In this section, we describe the proposed rectification method to transform horizontal and vertical birefractive baseline vectors into vertical baseline vectors only.

Traditional binocular stereo model formulates disparity as $r_{\text{binocular}}(P, z) = (f/z) b_{\text{binocular}}$, where $r_{\text{binocular}}$ is binocular disparity and $b_{\text{binocular}}$ is the binocular baseline between stereo cameras [10]. The birefractive stereo model from [1] also has a similar form, explaining the birefractive disparity $r_{\overline{oe}}$, the disparity between P_o and P_e, as follows:

$$r_{\overline{oe}}(P_o, z) = (f/z) b_{\overline{oe}}(P_o, z), \tag{6}$$

where $b_{\overline{oe}}$ is the birefractive baseline vector, defined as:

$$b_{\overline{oe}}(P_o, z) = b_{\overline{od}}(P_o) + b_{\overline{de}}\left(P_o + r_{\overline{od}}(P_o, z)\right). \tag{7}$$

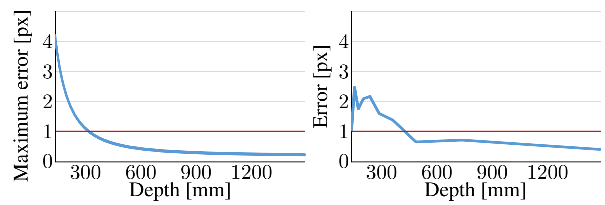
$b_{\overline{od}}$ and $b_{\overline{de}}$ are the baselines between P_o and P_d and between P_d and P_e. $r_{\overline{od}}$ is the disparity between P_o and P_d.

The key difference between the binocular and birefractive models is that $b_{\text{binocular}}$ in the binocular stereo model is *a constant* while $b_{\overline{oe}}$ in the birefractive stereo model changes depending on pixel position P_o and depth z. Owing to these two dependencies, to estimate depth per pixel, the current birefractive stereo model needs to estimate the birefractive baseline for every depth candidate and pixel position. Consequently, computation is expensive and has a large memory footprint. To overcome this limitation, we devise a novel rectification method for double refraction images that has no dependency on either depth or pixel position, enabling efficient birefractive stereo imaging with low memory footprint that is as fast as traditional binocular stereo imaging.

5.1. Depth Dependency of Birefractive Baseline

The birefractive baseline $b_{\overline{oe}}$ depends on both P_o and z, as shown in Equation (7). We evaluate the impact of P_o and z on the changes of $b_{\overline{oe}}$. We first found that the depth dependency of the baseline can be safely detached. Note that our goal is to derive a new disparity function $\hat{r}_{\overline{oe}}(P_o)$ with a depth-invariant baseline $\hat{b}_{\overline{oe}}(P_o)$ as follows:

$$\hat{r}_{\overline{oe}}(P_o, z) = (f/z) \hat{b}_{\overline{oe}}(P_o). \tag{8}$$



(a) Maximum approximation error (b) Error against Zemax

Figure 4. (a) We use Equation (8) to quantify errors induced by our approximation w.r.t. depth. (b) Our approximated model accurately predicts double refraction (measured at center), with results similar to those of a professional optics simulator, Zemax.

We found that, when depth z is larger than a specific value (410 mm in our optical setup (refer to Section 6 for details)), the depth dependency in the birefractive baseline of Equation (7) can be removed with errors of less than one pixel, resulting in the approximated baseline: $\hat{b}_{\overline{oe}}(P_o) = b_{\overline{od}}(P_o) + b_{\overline{de}}(P_o)$, which was used in our approximated birefractive stereo model in Equation (8). Refer to the supplementary document for our mathematical derivation details. Figure 4(a) shows that our approximated model is valid in terms of maximum error when $z > 410$ mm, and Figure 4(b) shows that our approximated model accurately simulates double refraction, with results similar to full optical ray tracing via Zemax.

5.2. Spatial Dependency of Birefractive Baseline

The approximated birefractive baseline $\hat{b}_{\overline{oe}}$ has no dependency on the depth, but it still depends on the spatial position of pixel P_o, resulting in spatially-varying magnitude and direction of the birefractive disparity $\hat{r}_{\overline{oe}}(P_o)$.

Here, we aim to detach the spatial dependency from the approximated birefractive baseline $\hat{b}_{\overline{oe}}(P_o)$ so that we can use line scans on the rectified input image to estimate the depth from the per-pixel refractive disparity and achieve *our final birefractive stereo model*, as follows:

$$\tilde{r}_{\overline{oe}}(z) = (f/z) \tilde{b}_{\overline{oe}}, \tag{9}$$

where $\tilde{r}_{\overline{oe}}$ is the birefractive disparity and $\tilde{b}_{\overline{oe}} = \left[\tilde{b}_{\overline{oe}}^{\text{avg}}, 0\right]$ is the birefractive baseline, whose horizontal and vertical components are set at $\tilde{b}_{\overline{oe}}^{\text{avg}}$ and zero, respectively. It is worth noting that $\tilde{b}_{\overline{oe}}^{\text{avg}}$ is a *constant scalar* as we set it as the average of $\hat{b}_{\overline{oe}}$ along the horizontal axis. Equation (9) now has a form with a constant baseline $\tilde{b}_{\overline{oe}}$, similar to that of the popular binocular stereo model. This change of the original spatially-varying baseline $\hat{b}_{\overline{oe}}(P_o)$ into the constant baseline $\tilde{b}_{\overline{oe}}$ causes the input image to follow the constant baseline setup via the ensuing rectification step.

Rectification via Dynamic Programming We introduce a novel rectification method that eliminates the spatial dependency of the birefractive baseline $\hat{b}_{\overline{oe}}(P_o)$ by warping the captured image. Our aim is to estimate a rectification func-

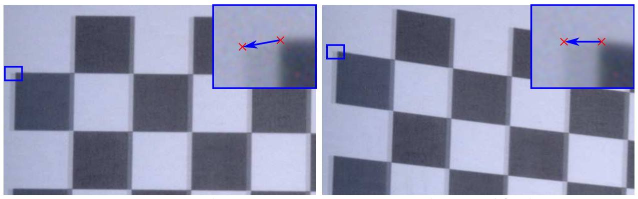


Figure 5. Spatial dependency of disparity and rectification. (a) Image captured before rectification exhibiting non-horizontal disparity. (b) Rectified image with horizontal and standardized disparity.

tion T that maps the input pixel P to the corresponding rectified pixel $\tilde{P}$: $P = T(\tilde{P})$. Once the function T is known, we can transform the input image I_c into the rectified version $\tilde{I}_c$: $\tilde{I}_c(\tilde{P}) = I_c(T(\tilde{P}))$.

The rectification function T is defined to make Equation (9) hold for the rectified image. To this end, we propose a dynamic programming algorithm, which defines T for each column from left to right. Following the principle of dynamic programming, we first initialize T for the first column in order for it to have an identical location before and after the rectification: $T([0, y]) = [0, y]$, where y is a row. As the second step of dynamic programming, we define T for a column x by assuming that T is known for the previous columns:

$$T(\tilde{P}) = T(\tilde{P} - [1,0]) + \hat{b}_{\overline{oe}}\left(T(\tilde{P} - [1,0])\right)/\tilde{b}^{\text{avg}}_{\overline{oe}}. \quad (10)$$

Equation (10) starts with the known T of the previous column and has an additional offset: $\hat{b}_{\overline{oe}}\left(T(\tilde{P} - [1,0])\right)/\tilde{b}^{\text{avg}}_{\overline{oe}}$. This offset simply maps the previous spatially-varying baseline $\hat{b}_{\overline{oe}}\left(T(\tilde{P} - [1,0])\right)$ to the target constant baseline $\tilde{b}^{\text{avg}}_{\overline{oe}}$. Therefore, it ensures that function T satisfies Equation (9). Figure 5 shows that our algorithm is able to warp a captured image with non-constant and non-horizontal disparity into a rectified image that satisfies both requirements.

6. Results

Hardware Implementation We built our experimental setup using a machine-vision camera (GS3-U3-123S6C-C) with the pixel pitch of $3.45\,\mu$m. For optical elements, we used a 35 mm lens, a glass-type linear polarizer from Edmund Optics, and a 15 mm thick calcite crystal from Newlight Photonics. Note that by increasing the thickness of the calcite, we can increase the disparity range in our method. The refractive indices of the calcite crystal are given as 1.65 and 1.48 for o-ray and e-ray, respectively, within the visible spectrum. To obtain deep depth-of-field, the aperture was set to $f/22$. The attenuation ratio τ between the o-ray and e-ray was calibrated by capturing stripe patterns and measuring the intensity ratios around the edges. Note that following Zhang [37] and Baek et al. [1], we also calibrated

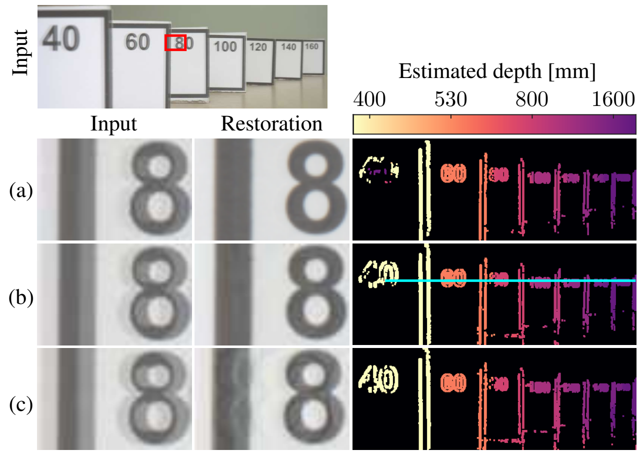


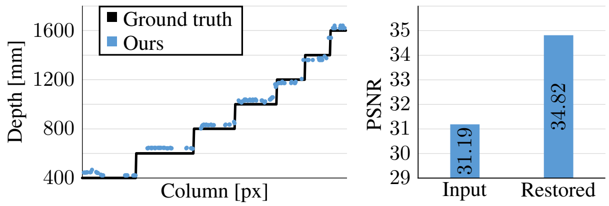
Figure 6. Our results with different polarizer orientations: (a) 0.15, (b) 0.3, (c) 0.45. Lower values of τ, (a) & (b), lead to clean restorations of the image, with fast convergence. The lowest e-ray proportion (a) leads to lower depth accuracy. Higher values of τ (c) result in image artifacts. For a reliable reconstruction of depth and color, we finally chose $\tau = 0.3$.

camera parameters and birefringent properties of the calcite crystal. Refer to the supplemental material for more details.

Software Implementation We implemented our main algorithm for joint depth and color reconstruction in *C++* using OpenCL GPU acceleration, while the birefractive model computation and the calibration process were written in MATLAB. We tested our reconstruction implementation on a computer configuration with an Intel core i7-7700K 4.2 GHz and an NVIDIA GTX 1080 Ti. For the image resolution of 2048×1500 and 16 depth candidates, our algorithm runs within 34 ms per each frame (30 Hz) for depth and color estimation. In details, rectification and restoration-volume generation take 16 ms. Cost computation and depth selection take 14 ms and 4 ms for computing validity mask.

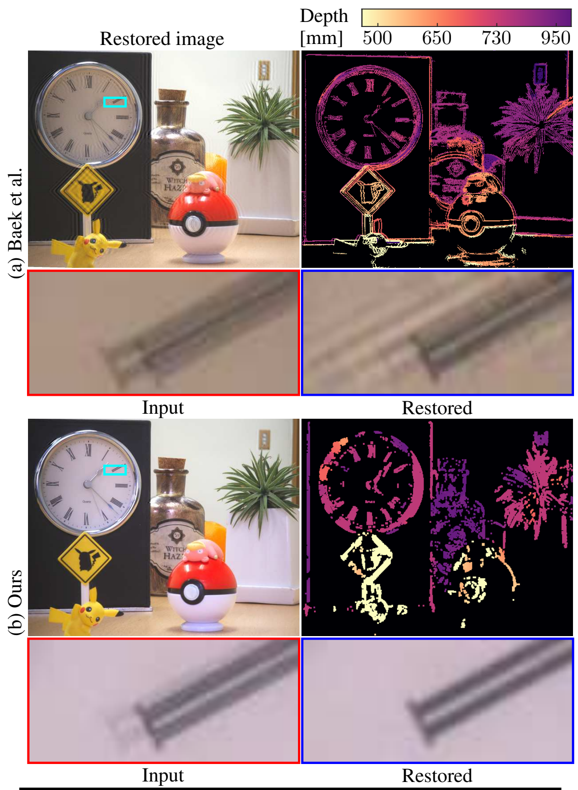
Unevenness of Double Refraction It is critical to determine the intensity proportion of e-ray to o-ray, τ; accurate determination of this value results in a clear reconstruction of image and depth. The residual error of our color restoration algorithm after N iterations is given by Equation 3. This demonstrates that the residual error is lower if τ is small. However, this holds only when the weak refraction clearly stands from the image noise. To determine the best value of τ, we captured a scene with panels (Figure 6). By adjusting the angle between the linear polarizer and the calcite crystal, we tested three different values of τ: 0.15, 0.3 and 0.45. We experimentally chose $\tau = 0.3$ and use it in all experiments.

Evaluation on Real Data For evaluation on real data with ground truth, we used the panel scene in Figure 6 with known panel distances. To validate the accuracy of our method, Figure 7(a) shows a 1D plot of depth estimates for



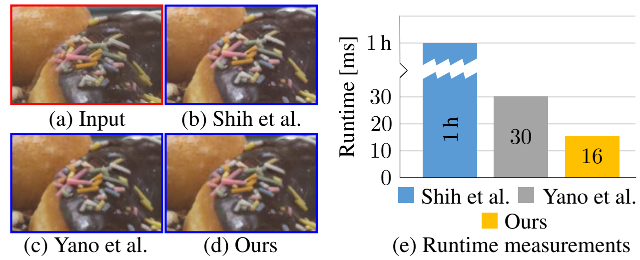
(a) Ours against ground truth (b) Restoration evaluation

Figure 7. (a) Quantitative depth values read along the light blue line of our depth map results (Figure 6(b)). (b) PSNR values of input and restoration w.r.t. reference.



	Baek (CPU)	Ours (CPU)	Ours (GPU)
Avg. runtime (ms)	38000	694	**34**
Max. memory (GB)	4.10	0.46	**0.46**
Input color PSNR (dB)	25.60	32.20	**32.20**
Output color PSNR (dB)	26.71	36.63	**36.63**
Depth RMSE (mm)	212	116	**116**

Figure 8. Comparison with the existing double refraction method. Baek's method (a) shows errors with the front clock and the back bottle in the depth estimate. The restored image severely suffers from ringing artifacts. Our method (b) can estimate depth more accurately, yielding a high-quality color image. The table compares computational time and accuracy of color and depth with synthetic ground truth. On the same CPU platform, our method is ~55-times faster than Baek's method. Our GPU implementation is ~20-times faster than our CPU version. Note that Baek's method is not GPU-friendly because it includes non-local cost aggregation [36] and dual cost computation [1].



(a) Input (b) Shih et al. (c) Yano et al. (d) Ours (e) Runtime measurements

Figure 9. Image restoration comparison. (a) Input. Images (b)–(d) are restored images by state-of-the-art methods [25, 33] and ours. They all are competitive in terms of image quality; however, the computational time is significantly different. It took the shortest time, just 16 ms, for our method to restore the color.

the panel scene, compared to the ground truth (measured by a Bosch GLM 80 laser meter). The averaged depth error of all the panels is 4.7 cm. Compared with the ground-truth photograph of the o-ray image, our restored image achieved a peak-signal-to-noise-ratio (PSNR) of 34.82 dB (Figure 7(b)), validating the effectiveness of our method. See Supplemental Material for additional image results.

Comparison with Depth-from-Double-Refraction We compared our method with the existing depth-from-double-refraction method [1], using the authors' original implementation. We achieved high accuracy on both color and depth estimates (Figure 8). While the previous method suffers from artifacts in color restoration (PSNR 26.71 dB), degraded depth quality (RMSE 212 mm) with high computational burden (runtime for depth 38 sec., runtime for color 78 sec., and memory footprint 4.1 GB), our method exploits the uneven double refraction with our joint reconstruction equipped with our rectification, outperforming the previous art by significant margins: color accuracy (PSNR 36.63 dB), depth accuracy (RMSE 116 mm), and computational efficiency (runtime 694/34 ms (CPU/GPU) and memory footprint (0.46 GB)). Refer to the supplemental material for more comparisons with a light-field camera, a dual-pixel camera, and a learned-based method [35].

Evaluation on Synthetic Data For evaluation with per-pixel ground truth, we created a synthetic dataset by simulating our image formation with 23 images of Middlebury dataset [22] with depth values between 400 and 1600 mm and inserted Gaussian noise of standard deviation 0.0005. The average PSNR of the restored color image is 36.63 dB and the average depth RMSE is 116 mm. Refer to the supplemental document for further qualitative and quantitative results on the dataset.

Ablation Study We ablate each component of our method to evaluate their respective impact on the performance using the same dataset on Table 1. Compared with [1], our novel rectification method reduces memory footprint and computational time significantly. Our optical design makes the color restoration problem much less ill-posed, which dramatically improves the color image quality.

Rectification	×	◦	◦	◦
Uneven double refraction	×	×	◦	◦
Joint reconstruction	×	×	×	◦
Avg. runtime ms	38000	27000	27000	**34**
Max. memory (GB)	4.10	2.70	2.70	**0.46**
Output color PSNR (dB)	26.71	26.80	33.23	**36.63**
Depth RMSE (mm)	212	193	445	**116**

Table 1. Averaged ablation study results with the synthetic dataset.

Comparison with Image Restoration Methods We compare our method on image restoration of uneven double refraction with those of existing deconvolution methods [25, 33]. While the restored image qualities are highly competitive, the computational costs are significantly different. It took just 16 ms for our method to restore the color image (see the table in Figures 9). We use the authors' implementation for Shih et al. [25] (written in Matlab); hence, speed is not directly comparable. We implemented Yano et al. [33] and ours using OpenCL for a fair comparison.

Applications Our method provides a sparse depth map and a restored image per each frame input enabling 3D object detection with the estimated sparse depth. We used FrustumNet v1 architecture [21] and retrained it for taking the sparse depth estimates of our method. To this end, we generated another synthetic dataset of 300 pairs of an uneven double-refraction image, a sparse depth map estimated by our method, and object labels. Specifically, we used 300 images of SUNRGBD dataset [27] captured by Kinect v2 devices, which provide high spatial and depth resolution. Note that the selected 300 images contain three object classes of table, desk, and chair mostly. We then simulate uneven double-refraction images from which we can estimate a sparse depth map assuming 30 mm thick calcite to handle the large depth range of the dataset. Figure 10 shows the detected 3D objects on test scenes. This experiment validates that our RGB-D output can be used successfully for the 3D object detection task. In Table 2, our mean average precision (mAP) value is highly competitive to the detection results trained with the full depth input.

We demonstrate three depth-aware image refocusing by densifying our sparse depth estimates guided by the restored RGB image, as shown in Figure 10. For densification, we used the fast bilateral solver [3], which runs in 70 ms, resulting in 104 ms for the full pipeline. Refer to the supplementary for details and other image editing applications.

7. Discussion and Conclusions

Our method is not free from limitations that can lead to interesting future work. Specifically, saturated regions and defocus and motion blur pose challenges for reconstruction. Future methods may rely on semantic feedback to the reconstruction algorithm to tackle these scenarios.

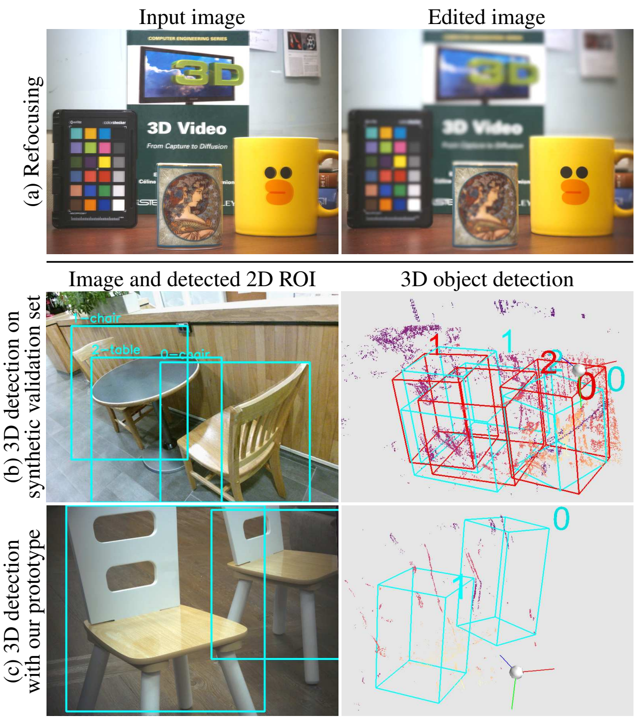


Figure 10. (a) Synthetic defocus on the background of the scene using our estimated color and depth. (b), (c) We trained the FrustumNet model [21] for our RGB-D camera. We use the SUNRGBD dataset [27] to generate synthetic birefractive images to train this model. (b) The synthetic result shows that our detection result has a good agreement with the ground truth. (c) The detection result captured by our real prototype. It detects 3D object volumes of chairs and their orientation successfully.

	Table	Desk	Chair	mAP
Trained with our sparse depth	0.79	0.73	0.67	0.74
Trained with dense depth (GT)	0.86	0.80	0.94	0.86

Table 2. 3D object detection AP trained with sparse depth maps and the ground truth dense depth, using the AP metric [21].

We have presented a real-time monocular RGB-D imaging method relying on uneven double refraction, i.e., the cross-polarization property of double refraction. The proposed joint depth and color reconstruction method efficiently and accurately estimates sparse depth and dense color, outperforming previous depth-from-double refraction methods in accuracy, while being orders of magnitudes faster. We have validated the proposed method both synthetically and experimentally, and demonstrate 3D object detection and photographic applications.

Acknowledgements

Min H. Kim acknowledges Korea NRF grants (2019R1A2C3007229, 2013M3A6A6073718), KOCCA in MCST of Korea, Samsung Research, and Cross-Ministry Giga KOREA (GK17P0200).

References

[1] Seung-Hwan Baek, Diego Gutierrez, and Min H Kim. Birefractive stereo imaging for single-shot depth acquisition. *ACM Trans. Graphics (Proc. SIGGRAPH Asia 2016)*, 35(6):194, 2016.

[2] Yosuke Bando, Bing-Yu Chen, and Tomoyuki Nishita. Extracting depth and matte using a color-filtered aperture. *ACM Trans. Graphics (Proc. SIGGRAPH Asia 2008)*, 27(5):134, 2008.

[3] Jonathan T Barron and Ben Poole. The fast bilateral solver. *Proc. European Conference on Computer Vision (ECCV) 2016*, 2016.

[4] Zhihu Chen, Kwan-Yee K Wong, Yasuyuki Matsushita, and Xiaolong Zhu. Depth from refraction using a transparent medium with unknown pose and refractive index. *International Journal of Computer Vision*, 102(1-3):3–17, 2013.

[5] John Ens and Peter Lawrence. An investigation of methods for determining depth from focus. *IEEE Transactions on pattern analysis and machine intelligence*, 15(2):97–108, 1993.

[6] Chunyu Gao and Narendra Ahuja. Single camera stereo using planar parallel plate. In *Proc. Int. Conf. Pattern Recognition (ICPR)*, volume 4, pages 108–111, 2004.

[7] Chunyu Gao and Narendra Ahuja. A refractive camera for acquiring stereo and super-resolution images. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 2316–2323, 2006.

[8] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T. Barron. Learning single camera depth estimation using dual-pixels, 2019.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, volume 2, page 7, 2017.

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. 2003.

[11] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Experimental robotics*, pages 477–491. Springer, 2014.

[12] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 35(2):504–511, 2013.

[13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[14] Kurt Konolige. Projected texture stereo. In *2010 IEEE International Conference on Robotics and Automation*, pages 148–155. IEEE, 2010.

[15] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000.

[16] DooHyun Lee and InSo Kweon. A novel stereo camera system by a biprism. *IEEE Trans. Robotics and Automation*, 16(5):528–541, 2000.

[17] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graphics (Proc. SIGGRAPH 2007)*, 26(3):70, 2007.

[18] Larry Matthies, Richard Szeliski, and Takeo Kanade. Incremental estimation of dense depth maps from image sequences. In *Proceedings CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition*, pages 366–374. IEEE, 1988.

[19] Julia Navarro and Antoni Buades. Robust and dense depth estimation for light field images. *IEEE Transactions on Image Processing*, 26(4):1873–1886, 2017.

[20] Y Nishimoto and Y Shirai. A feature-based stereo model using small disparities. In *Proc. Comput. Vision and Pattern Recognition (CVPR)*, pages 192–196, 1987.

[21] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

[22] Daniel Scharstein, Heiko Hirschmuller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proc. German Conf. Pattern Recognition*, pages 31–42. Springer, 2014.

[23] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. volume 1, 2003.

[24] Jianping Shi, Xin Tao, Xu Li, and Jiaya Jia. Break ames room illusion: Depth from general single images. *ACM Trans. Graphics (Proc. SIGGRAPH Asia 2015)*, 2015.

[25] YiChang Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3201, June 2015.

[26] Masao Shimizu, Masatoshi Okutomi, and Wei Jiang. Disparity estimation in a layered image for reflection stereo. In *Proc. Asian Conference on Computer Vision (ACCV) 2009*, pages 395–405, 09 2009.

[27] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, June 2015.

[28] Murali Subbarao and Gopal Surya. Depth from defocus: a spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, 1994.

[29] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM Trans. graphics (TOG)*, volume 26, page 69. ACM, 2007.

[30] Neal Wadhwa, Marc Levoy, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, and Yael Pritch. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics*, 37(4):113, Jul 2018.

[31] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495, 2015.

[32] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Depth estimation with occlusion modeling using light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2170–2181, 2016.

[33] T. Yano, M. Shimizu, and M. Okutomi. Image restoration and disparity estimation from an uncalibrated multi-layered image. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) 2010*, pages 247–254, 2010.

[34] Jinwei Ye, Yu Ji, Wei Yang, and Jingyi Yu. Depth-of-field and coded aperture imaging on xslit lens. In *Proc. European Conference on Computer Vision (ECCV)*, pages 753–766, 2014.

[35] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[36] Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, Shuicheng Yan, and Qi Tian. Cross-scale cost aggregation for stereo matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014.

[37] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22, 2000.