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

Andreas Meuleman\textsuperscript{1*} Seung-Hwan Baek\textsuperscript{1* †} Felix Heide\textsuperscript{2} Min H. Kim\textsuperscript{1}

\textsuperscript{1}KAIST  \textsuperscript{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.

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

\textsuperscript{*}: Equal contribution. \textsuperscript{†}: 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 \( P_o \) 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 \( r_{\text{m}} \) 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 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 \( \tilde{r}_{\text{m}}(z) \). 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 \( \tau \), obtained through calibration (refer to the supplemental document for more details). Therefore, the e-ray image \( \tilde{I}_e \) can be formulated as: \( \tilde{I}_e = \tau A(\tilde{I}_o, \tilde{r}_{\text{m}}(z)) \), where \( A(\tilde{I}_o, \tilde{r}_{\text{m}}(z)) \) is a function that translates the o-ray image \( \tilde{I}_o \) according to the disparity \( \tilde{r}_{\text{m}}(z) \). 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:

\[
\tilde{I}_c = \tilde{I}_o + \tau A(\tilde{I}_o, \tilde{r}_{\text{m}}(z)) .
\] (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 2. Our image formation for uneven double refraction. A scene point \( P_o \) is directly projected to \( P_d \), while o-ray and e-ray of \( P_o \) 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 \( r_{\text{m}} \).](image-url)
The key idea of our image restoration is to iteratively remove the image residuals, but instead introduce false edges as artifacts. Therefore, we define the depth cost volume for every 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 $\hat{I}_o$.

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

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

As the attenuation ratio of e-ray $\tau$ is by definition less than one, the new residual $\tilde{\Delta} z^{(1)} = -\tau^2 A(\tilde{\hat{I}}_o, 2\hat{r}_\text{pp}(z))$ has a lower intensity level than that of the previous residual $\Delta z^{(0)}$, making our current estimate $\tilde{\hat{I}}_o^{(1)}$ closer to the ground truth than the previous estimate $\tilde{\hat{I}}_o^{(0)}$. In the next iteration, the approximated residual is similarly defined as follows: $\tilde{\Delta} z^{(1)} = -\tau^2 A(\tilde{\hat{I}}_o^{(1)}, 2\hat{r}_\text{pp}(z)) = -\tau^2 A(\tilde{\hat{I}}_o, 2\hat{r}_\text{pp}(z)) + \tau^4 A(\tilde{\hat{I}}_o, 4\hat{r}_\text{pp}(z))$. The current image estimate is then also updated as: $\hat{I}_o^{(2)} = \hat{I}_o^{(1)} - \Delta z^{(1)} = \hat{I}_o + \tau^4 A(\hat{I}_o, 4\hat{r}_\text{pp}(z))$. 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

$$\Delta z^{(N)} = -\tau^{2N} A(\hat{I}_o, 2^{N} \cdot \hat{r}_\text{pp}(z)).$$

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

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

where $K(P)$ is the set of pixels in a window centered at $P$ of size $61 \times 61$. 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:

$$Z(P) = \arg \min_z C^z(P).$$
With the estimated depth $Z$, we can reconstruct the final color image $I_o^Z$ from the restoration volume $I_o$ (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 ($|\partial I_o^Z / \partial x| > \text{Thres}_{\text{grad}}$) 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 $b_{\text{binocular}}$ 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 $r_{\text{bire}}$, the disparity between $P_o$ and $P_e$, as follows:

$$r_{\text{bire}}(P_o, z) = (f/z)b_{\text{bire}}(P_o, z), \quad (6)$$

where $b_{\text{bire}}$ is the birefractive baseline vector, defined as:

$$b_{\text{bire}}(P_o, z) = b_{\text{od}}(P_o) + b_{\text{oe}}(P_o + r_{\text{od}}(P_o, z)). \quad (7)$$

$b_{\text{od}}$ and $b_{\text{oe}}$ are the baselines between $P_o$ and $P_d$ and between $P_d$ and $P_e$. $r_{\text{od}}$ 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 $b_{\text{bire}}$ 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 $b_{\text{bire}}$ 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 $b_{\text{bire}}$. 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 $\hat{r}_{\text{bire}}(P_o)$ with a depth-invariant baseline $\hat{b}_{\text{bire}}(P_o)$ as follows:

$$\hat{r}_{\text{bire}}(P_o, z) = (f/z)\hat{b}_{\text{bire}}(P_o) \quad (8)$$

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: $\hat{b}_{\text{bire}}(P_o) = b_{\text{od}}(P_o) + b_{\text{oe}}(P_o)$, 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 $\hat{b}_{\text{bire}}$ 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 $\hat{r}_{\text{bire}}(P_o)$.

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

$$\tilde{r}_{\text{bire}}(z) = \hat{r}_{\text{bire}}(z)\hat{b}_{\text{bire}}, \quad (9)$$

where $\tilde{r}_{\text{bire}}$ is the birefractive disparity and $\hat{b}_{\text{bire}} = \begin{bmatrix} \hat{b}_{\text{bire}}^{\text{avg}} & 0 \end{bmatrix}$ is the birefractive baseline, whose horizontal and vertical components are set at $\hat{b}_{\text{bire}}^{\text{avg}}$ and zero, respectively. It is worth noting that $\hat{b}_{\text{bire}}^{\text{avg}}$ is a constant scalar as we set it as the average of $b_{\text{bire}}$ along the horizontal axis. Equation (9) now has a form with a constant baseline $\hat{b}_{\text{bire}}$, similar to that of the popular binocular stereo model. This change of the original spatially-varying baseline $b_{\text{bire}}(P_o)$ into the constant baseline $\hat{b}_{\text{bire}}$ 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 $b_{\text{bire}}(P_o)$ by warping the captured image. Our aim is to estimate a rectification func-
Following Zhang [37] and Baek et al. [1], we also calibrated the intensity ratios around the edges. Note that for e-ray was calibrated by capturing stripe patterns and measuring the intensity ratios around the edges. To determine the best value of $\tau$, 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 $\tau$: 0.15, 0.3, and 0.45. We experimentally chose $\tau = 0.3$.

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 $2048 \times 1500$ 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. $\tau$: 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 $\tau$ is small. However, this holds only when the weak refraction clearly stands from the image noise. To determine the best value of $\tau$, 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 $\tau$: 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
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 dra-
matically 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.

**Acknowledgements**

References


