High-Quality Stereo Image Restoration from Double Refraction

Hakyeong Kim Andreas Meuleman Daniel S. Jeon Min H. Kim

KAIST

Abstract

Single-shot monocular birefractive stereo methods have been used for estimating sparse depth from double refraction over edges. They also obtain an ordinary-ray (o-ray) image concurrently or subsequently through additional post-processing of depth densification and deconvolution. However, when an extraordinary-ray (e-ray) image is restored to acquire stereo images, the existing methods suffer from very severe restoration artifacts due to a low signal-to-noise ratio of input e-ray image or depth/deconvolution errors. In this work, we present a novel stereo image restoration network that can restore stereo images directly from a double-refraction image. First, we built a physically faithful birefractive stereo imaging dataset by simulating the double refraction phenomenon with existing RGB-D datasets. Second, we formulated a joint stereo restoration problem that accounts for not only geometric relation between o-/e-ray images but also joint optimization of restoring both stereo images. We trained our model with our birefractive image dataset in an end-to-end manner. Our model restores high-quality stereo images directly from double refraction in real-time, enabling high-quality stereo video using a monocular camera. Our method also allows us to estimate dense depth maps from stereo images using a conventional stereo method. We evaluate the performance of our method experimentally and synthetically with the ground truth. Results validate that our stereo image restoration network outperforms the existing methods with high accuracy. We demonstrate several image-editing applications using our high-quality stereo images and dense depth maps.

1. Introduction

Double refraction occurs by birefringence, an optical property of anisotropic, transmissive materials, where an incident ray is split into two rays: ordinary ray (o-ray) and extraordinary ray (e-ray). A double-refraction image is a superimposed image caused by the refracted o-ray and e-ray placed on the same image with displacement (see Figure 1 for an example). Displacement in a double-refraction image contains additional information related to depth, similar to a disparity in traditional stereo, i.e., it is inversely proportional to depth. Based on this phenomenon, single-shot monocular birefractive stereo methods [1, 12] have been developed to estimate sparse depth from double refraction over edges, achieving passive monocular 3D imaging with a small form factor.

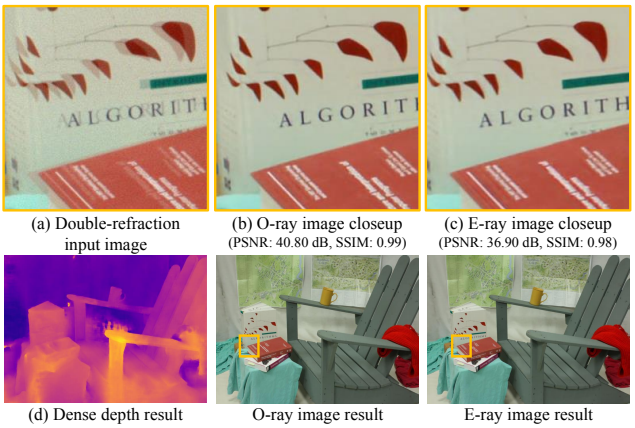


Figure 1: (a) an input double-refraction image, (b) an o-ray image result, (c) an e-ray result, (d) a dense disparity map estimated from our stereo images. Compared to ground truth, PSNRs and SSIMs of the entire o-/e-ray images are 40.80/36.90 dB, SSIM: 0.9854/0.9799, respectively.

To obtain two stereo images from double refraction, there are two different ways in the existing methods. One way is to densify the sparse depth map with a diffusion method [10] and then deconvolve the input double-refraction image with different point spread functions (PSFs) for each depth [1]. The resulting quality depends on the accuracy of per-pixel depth. Ringing artifacts of deconvolution often occur, degrading the image quality of the restored images. Another way is to subtract the restored o-ray image from the input double-refraction image and then recover its scale [12]. This approach restores an e-ray image from uneven double refraction. Still, the signal-to-noise ratio (SNR) of the input e-ray image is inherently low in an uneven double refraction image [12]. Even though the restored o-ray image is sharp, the e-ray image suffers from

severe noise, and also some restoration errors in the o-ray image are inherited from the e-ray image. Earlier birefractive imaging solutions mainly focus on acquiring the depth information and the ordinary-ray image only. None of the existing works can restore stereo images from a double-refraction image with high quality yet.

Different from the existing works, we focus on *stereo image restoration* directly from double refraction *without relying on depth information*. Since double refraction and displacement are geometrically related, our objective is challenging. Therefore, to mitigate the ill-posedness of our problem, we first generate a physically faithful birefractive stereo dataset from existing RGB-D datasets [14, 17] by simulating the double refraction phenomenon. Second, we formulate a joint stereo restoration problem that accounts for the geometric relation of o-ray and e-ray images to jointly infers both stereo images from double refraction with high accuracy. We train our model with our birefractive image dataset with supervised signals in an end-to-end manner. Our model can restore high-quality stereo images directly from double refraction (Figure 1a) without knowing depth information (Figures 1b and 1c). These stereo images also agree with epipolar geometry with high accuracy, allowing us to estimate dense depth with high accuracy from the restored stereo images using a conventional stereo method [3] (Figure 1d).

Our model is computationally efficient. It takes just $\sim$227 ms to restore two stereo images of three megapixels from input on a conventional desktop computer with a GPU. Using our network, a high-quality stereo video can be obtained directly from a double-refraction video in real-time, enabling high-quality anaglyph of 3D stereo vision using a single monocular camera.

We evaluate the performance of our method experimentally and synthetically with the ground truth. Results validate that our stereo image restoration network outperforms the existing methods with high accuracy. We demonstrate several image-editing applications using high-quality stereo images and dense depth maps. All codes and models are published to ensure reproducibility.

2. Related Work

Birefractive Stereo. Baek et al. [1] proposed a birefractive depth acquisition method that consists of a conventional camera and a calcite crystal attached in front of the camera lens. Double refraction by the birefringent material splits the ray into two rays: o-ray and e-ray. Therefore, the birefractive stereo system's captured image is a double-layered image of the identical scene but shifted by the spatially variant disparity, which is inversely proportional to the scene depth. Baek et al. [1] proposed an image formation model of birefractive stereo and analytically computed the disparity between o-ray and e-ray to depth. Meuleman et al. [12] introduced real-time birefractive stereo imaging by attaching a linear polarizer in front of the crystal to weaken one of the polarized rays intentionally. Uneven double refraction is beneficial to relax the ambiguity of correspondence search. This work restores o-ray images with improved image quality. Based on the analysis of the relation between the attenuation coefficients of double-refraction images and the restoration quality, we also decided to use uneven double refraction as an input of our stereo restoration network.

Image Restoration from Double Refraction. In the previous birefractive stereo methods, Baek et al. [1] recovered a clean o-ray image from a double-refraction image by formulating the problem into non-blind deconvolution. The pre-calibrated deconvolution kernel is voted using the estimated depth values. Oppositely, Meuleman et al. [12] restore the image in precedence to select the most probable disparity. In this work, we separate an overlapped two rays image into two stereo images of o-ray and e-ray without having depth information using a neural network.

Learning-based Image Restoration. In recent years, many single image non-uniform blind deblurring algorithms have been developed by employing deep learning methodologies. Instead of estimating blur kernels per image regions and combining with non-blind deconvolution, these studies have used a convolutional neural network (CNN) to directly reconstruct sharp latent images from blurry input images with high accuracy. Network architectures used for these methods includes multi-scale CNN [13], formatted residual network [5], and recurrent network [18]. While letting CNN as an end-to-end pipeline that generates latent images, studies have diversified by adopting adversarial learning [8] and supplementing a deconvolutional module that consists of recurrent neural network (RNN) [21] or deformable convolution modules [20]. However, existing works seek only a single latent image and discard residuals caused by blur.

Some recent studies related to superimposed image separation restore two latent images from mixed images, but their input data are grounded to the reflection of background scene on a transmitted foreground scene on glass [6, 22] or mixed input of completely different scenes [4, 24]. These recent works assume that the restoration target images must have distinctive differences in image statistics. However, for double refraction images, o-ray and e-ray images are very similar to each other with subtle displacement. Due to the statistical ambiguity in the superimposed double-refraction, images of o-/e-ray cannot be restored clearly using state-of-the-art restoration networks. Instead, our method leverages the resemblance with consideration of their geometric relation of o-/e-ray.

To the best of our knowledge, there is no network architecture that separates and restores both the latent image and

the derivative (or by-product) from the degradation of the equivalent scene in practical applications. Therefore, we propose a novel network architecture that differs from conventional image restoration networks to restore two stereo images of o-ray and e-ray efficiently.

3. Stereo Restoration from Double Refraction

3.1. Double Refraction

Image Formation Model. Birefringent material has an anisotropic optical property determined by its optical axis; it splits the light propagating directions into two by polarization. If an incoming ray has perpendicular polarization with the optical axis, the ray follows Snell's law and travels along the plane of incidence. This ray is called an ordinary ray (o-ray). Extraordinary ray (e-ray), the other ray among split rays, has parallel polarization with the optical axis. It propagates in the direction that walks off the plane of incidence, as illustrated in Figure 2(a). Therefore, each image generated by an o-ray and e-ray is the shifted image of the equivalent latent scene. Accordingly, the resulting double refraction image is in the form of overlapped underlying scenes as shown in Figure 2(b).

The early-stage depth acquisition method from double refraction [1] employs even double refraction images. However, finding correspondence points in even double refraction has to resolve ambiguity between o-ray and e-ray, doubling the computational costs. The subsequent study [12] resolves this issue by attaching the linear polarizer in front of the birefringent material. Since o-ray and e-ray are perpendicular in polarization direction, tuning the polarity of incoming light can adjust the intensity ratio of o-ray and e-ray. When e-ray is attenuated by the attenuation ratio of τ with respect to direct ray, uneven double refraction image $\mathbf{I}_b$ is formulated as following equation:

$$\mathbf{I}_b = \frac{\tau}{1+\tau}\mathbf{I}_e + \frac{1}{1+\tau}\mathbf{I}_o + \mu, \tag{1}$$

where $\mathbf{I}_e$, $\mathbf{I}_o$ and μ refers to e-ray, o-ray images and noise.

This work adopts this uneven double refraction model for stereo image acquisition under consideration of computational cost and stability of image restoration. The former reason coincides with the objective of the previous work [12] and the latter reason comes from the problem interpretation in the aspect of image deconvolution. Double refraction image can be modeled with image convolution with spatially varying kernel k, which is dependent on disparity between o-ray and e-ray d_{oe} with the following equation: $\mathbf{I}_b = k\left(d_{oe}\right) * \mathbf{I}_o + \mu$. The image formulation kernel assuming the rectification is depicted in Figure 2(c).

Since e-ray can be obtained from a spatially varying shift transformation of the o-ray image, the convolution kernel that simulates double refraction consists of two Dirac delta functions that each sample the latent signals with displacement. The intensity of kernel weights is determined by the attenuation ratio τ.

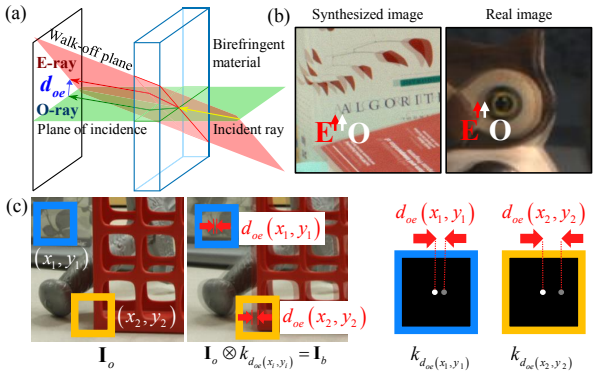


Figure 2: (a) Image formation by double refraction. (b) Uneven double refraction images; (left) rendered image with an attenuation ratio τ=0.3, (right) real double-refraction image captured with a birefractive stereo camera. (c) Interpretation of double-refraction image formation as convolution.

A previous study on image restoration from reflected images [19] provides an analysis of the stability of the deconvolution filter for two-layered images, whose generation kernel appears similar to the kernel of Figure 2(c). They examined the stability of their method on the varying intensity of a layered image and found that image restoration is stable when one of the layered image intensity is weakened. As such, we consider uneven double refraction is suitable for stable estimation of stereo images from uneven double refraction images.

Depth from Birefractive Disparity. The disparity d_{oe} between o-ray and e-ray is related to depth. While Baek et al. [1] maps disparity to depth z with the intermediate of a hypothetical *direct ray* – a ray corresponding to an absence of birefractive medium –, Meuleman et al. [12] have demonstrated that a mapping from disparity to depth can be achieved solely using disparity without significant approximation. In addition, they propose a rectification that reduces the disparity-to-depth conversion to: $z = \frac{c}{d_{oe}}$, where c is a baseline constant, similarly to conventional binocular stereo.

3.2. Stereo Restoration Network

Despite the adversity on restoring attenuated rays from a super-imposed image, a recent study on visually imbalanced stereo matching [11] addresses the importance of consistent and good visual quality of a stereo image pair for ideal depth estimation. Meuleman et al. [12] report a lower value of an attenuation ratio enhances image restoration quality when restoring a strong latent signal. Alternatively, we can induce that restoring the attenuated ray is analytically difficult.

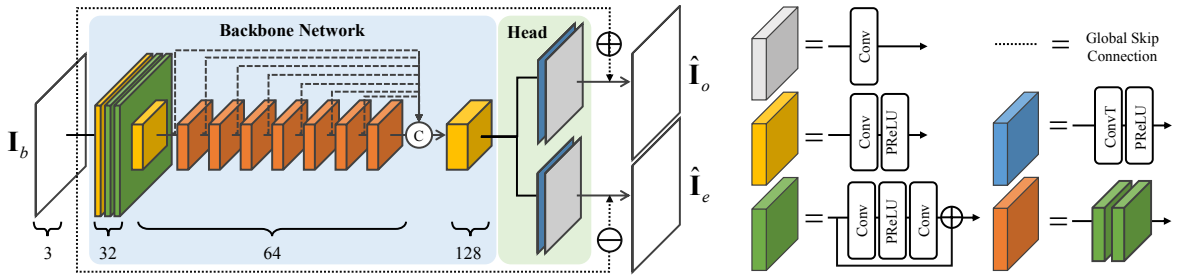


Figure 3: Architecture of our stereo restoration network. Refer to Section 3.2 for description of the network architecture.

To efficiently restore both o-ray and e-ray in competitive quality, we analyzed the residuals between the individual ray and double refraction signal. Derived from Equation (1), Equation (2) shows that subtraction of each ray signal from double refraction signal is proportional to the term $\mathbf{I_o} - \mathbf{I_e}$ when noise term is ignored as follows:

$$\mathbf{I}_b - \mathbf{I}_o = \tfrac{\tau}{1+\tau}\left(\mathbf{I}_e - \mathbf{I}_o\right), \quad \mathbf{I}_b - \mathbf{I}_e = \tfrac{1}{1+\tau}\left(\mathbf{I}_o - \mathbf{I}_e\right). \tag{2}$$

We design a novel stereo restoration network from the idea that successful feature extraction of the common term in Equation (2) with the supervision of strong latent signals would also aid the estimation of attenuated signals.

Network Architecture. Based on the motivation, we designed a stereo restoration network to have one backbone network, two head networks, and two global skip connections, as shown in Figure 3.

To train network layers to learn residuals between individual rays and double refraction, we adopt global skip connections for both head networks. Several studies have utilized global skip connections for performance enhancement in image deblurring, and restoration [20, 8]. We take advantage of global skip connection to effectively utilize our insight that the residual between o-ray and birefractive image is negatively proportional to the residual between e-ray and birefractive image. We design two types of global skip connections: one that element-wisely subtracts layer output from the input image, and the other that element-wisely adds layer output to the input image. These operations are summarized as:

$$\begin{aligned} \mathbf{I}_b - g\left(\mathbf{I}_b\right) &= \mathbf{I}_e, \\ \mathbf{I}_b + \tilde{g}\left(\mathbf{I}_b\right) &= \mathbf{I}_o, \end{aligned} \tag{3}$$

where g and $\tilde{g}$ denotes the forward operation of the backbone network and each head networks. Since layers that estimate residual of o-ray and layers that estimate residual of e-ray shares majority of parameters by using common backbone network, both rays are jointly estimated, and negative proportionality of two CNN outputs are induced by two different types of global skip connections.

Figure 3 illustrates the details of the network architecture. All convolutional layers in the network have a kernel size three and parametric ReLU layers for activation. The backbone network consists of a sequence of residual blocks. This architecture is inspired by the feature extractor of Yuan et al. [20], which uses stride convolution for downsampling instead of pooling layers. However, we do not use dilation to densely participate in neighboring pixels in feature computation while stacking more residual blocks to retain the size of receptive fields.

The head networks upsample image resolution with transposed convolution. As g and $\tilde{g}$ learn following quantities,

$$\begin{aligned} g\left(\mathbf{I}_b\right) &= \tfrac{1}{1+\tau}\left(\mathbf{I}_o - \mathbf{I}_e\right) + \mu, \\ \tilde{g}\left(\mathbf{I}_b\right) &= -\tfrac{\tau}{1+\tau}\left(\mathbf{I}_o - \mathbf{I}_e\right) - \mu, \end{aligned} \tag{4}$$

we allow bias in the transposed convolution layer and add a 1×1 convolution layer as the final layer.

Loss. For prevalent usage of l_2 loss in image restoration and regression as loss metric, we adopt l_2 loss defined as $\mathcal{L}_r(\mathbf{I}, \hat{\mathbf{I}}) = \|\mathbf{I} - \hat{\mathbf{I}}\|_2^2$, where $\mathbf{I}$ and $\hat{\mathbf{I}}$ are input and restored images. However, due to the limit of l_2 loss in reconstructing images' structural details, we fine-tune the network through further training with different loss metrics after training with l_2 loss. As addressed in [23], we use mixture of l_1 loss and MS-SSIM (multi-scale structural smilarity) loss $\mathcal{L}_f$ for better restoration quality. The loss metric is defined as:

$$\mathcal{L}_r\left(\mathbf{I}, \hat{\mathbf{I}}\right) = (1-\alpha)\left\|\mathbf{I} - \hat{\mathbf{I}}\right\|_1 + \alpha \cdot \mathcal{L}_f\left(\mathbf{I}, \hat{\mathbf{I}}\right), \tag{5}$$

where the ratio is set as $\alpha = 0.84$. We compute restoration loss between the restored ray and ground truth for both o-ray and e-ray. The total loss is the sum of restoration loss for each ray, formulated as: $\mathcal{L}_{total} = \lambda_o \mathcal{L}_r(\mathbf{I}_o, \hat{\mathbf{I}}_o) + \lambda_e \mathcal{L}_r(\mathbf{I}_e, \hat{\mathbf{I}}_e)$, where default ratios of λ_o, λ_e are set as 0.5.

Training/Test Datasets. Dataset to train and test our stereo restoration network should contain compilation of double refraction image $\mathbf{I}_b$, e-ray $\mathbf{I}_e$, o-ray $\mathbf{I}_o$ and disparity d_{oe} between o-ray and e-ray: $D = \{\mathbf{I}_b, \mathbf{I}_o, \mathbf{I}_e, d_{oe}\}$. Since any training dataset for birefractive stereo imaging does not exist, we synthesized a dataset with publicly available RGB-D datasets. Refer to Section 1 in the supplemental document for the technical details on how to simulate the training image dataset.

We set the attenuation ratio τ to 0.3 on the basis of analysis on polarizer orientation by [12]. Lastly, we added gaus-

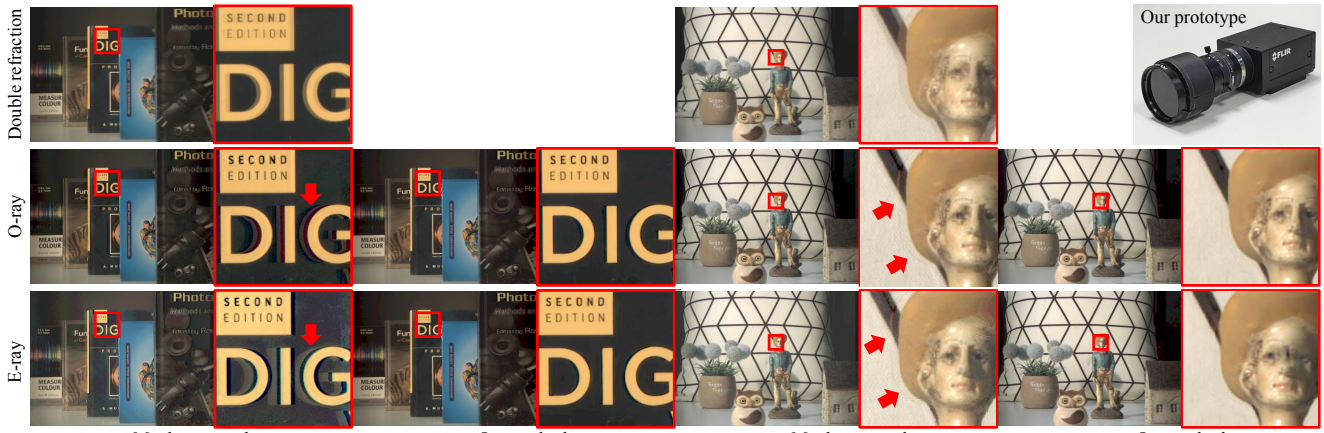


Figure 4: Comparison of our image restoration results on real double refraction images captured by our birefractive stereo camera with a birefractive stereo method [12], which used the same input of uneven double refraction. Our method can provide clearer restoration results than the existing method without suffering from ringing artifacts. Inset (top right) Our real-system prototype of birefractive stereo imaging.

sian noise with a standard deviation of 0.0005 on the synthesized double refraction images.

We adopted the labeled NYU-Depth v2 dataset [14] for generating train samples since it provides 1449 scenes with a dense depth map. We converted given depth in meters into disparity ranging between 0 to 32. Each 640×480 size images were sliced into 256×256 size patches for training.

We generate a test dataset with the Middlebury dataset 2014 [17]. It provides 23 high-resolution stereo image pairs of indoor scenes with ground truth disparity. We downsample images into size 2048×1500, and warped depth from stereo disparity to disparity between o-ray and e-ray. The warped disparity ranges from 5 to 20.

Implementation Details. Our stereo restoration network was implemented with Pytorch. We trained the network using a synthetic train dataset and Adam optimizer [7] on an NVIDIA Titan RTX GPU. We went through three phases of training to get the final model. To start, we set optimizer parameter as $\beta_1 = 0.9$ and $\beta_2 = 0.9$ and trained network with l_2 loss for first 100 epochs. The learning rate and batch size were set as 10^{-3} and 32. Then, we changed Adam parameters into $(\beta_1 = 0.9, \beta_2 = 0.999)$ and used AMSGrad [15] gradient direction. Network was trained further for 16 epochs with mixture of l_1 and MS-SSIM defined in Equation (5). Initial learning rate was 10^{-4} and it was halved for every 5 epochs. For the last phase, we changed the weight on e-ray restoration loss and o-ray restoration loss to $\lambda_e = 0.75$ and $\lambda_o = 0.25$ for further enhancement on e-ray restoration. We let Adam optimizer have default parameter settings with a learning rate of 10^{-4}. We applied the weight decaying parameter as 10^{-5} in this phase.

4. Results

Hardware Prototype. We built a prototype of birefractive stereo imaging with uneven double refraction and calibrated it as proposed by Meuleman et al. [12] (Figure 4 inset). We employed a machine-vision camera (GS3-U3-123S6C-C) with a 35 mm lens, a glass-type linear polarizer from Edmund Optics, and a 15 mm thick calcite crystal from Newlight Photonics. The refractive indices of the crystal are 1:65 and 1:48 for o-ray and e-ray, respectively. To obtain deep depth-of-field, the aperture was set to $f/22$.

Figure 4 presents stereo restoration results from double-refraction images captured by our real prototype. We compare our results with restored o-/e-ray images by Meuleman et al. [12]. This figure shows that our method restores stereo images in clear shape and retains the displacement between o-ray and e-ray simultaneously. On the other hand, we can observe that artifacts in the o-ray restoration of the other method [12] have a direct influence on the failure of e-ray image restoration.

Quantitative Evaluation. We compare the accuracy of our stereo image restoration method with two birefractive stereo methods [1, 12] on unseen synthetic double refraction scenes generated from a Middlebury RGB-D dataset [17]. Note that per-pixel ground truths for both o-ray and e-ray are available only in this synthesized dataset. Since earlier studies do not produce e-ray images, we compute e-ray signals by subtracting the restored o-ray signals from the superimposed double refraction image and rescaling it (Equation (1)) as: $\mathbf{I}_e = \frac{1}{\tau}\left((1+\tau)\,\mathbf{I}_b - \mathbf{I}_o\right)$.

Table 1 summarizes the average peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) values of the restored o-ray and e-ray of each method. The results show that our method has significantly improved the

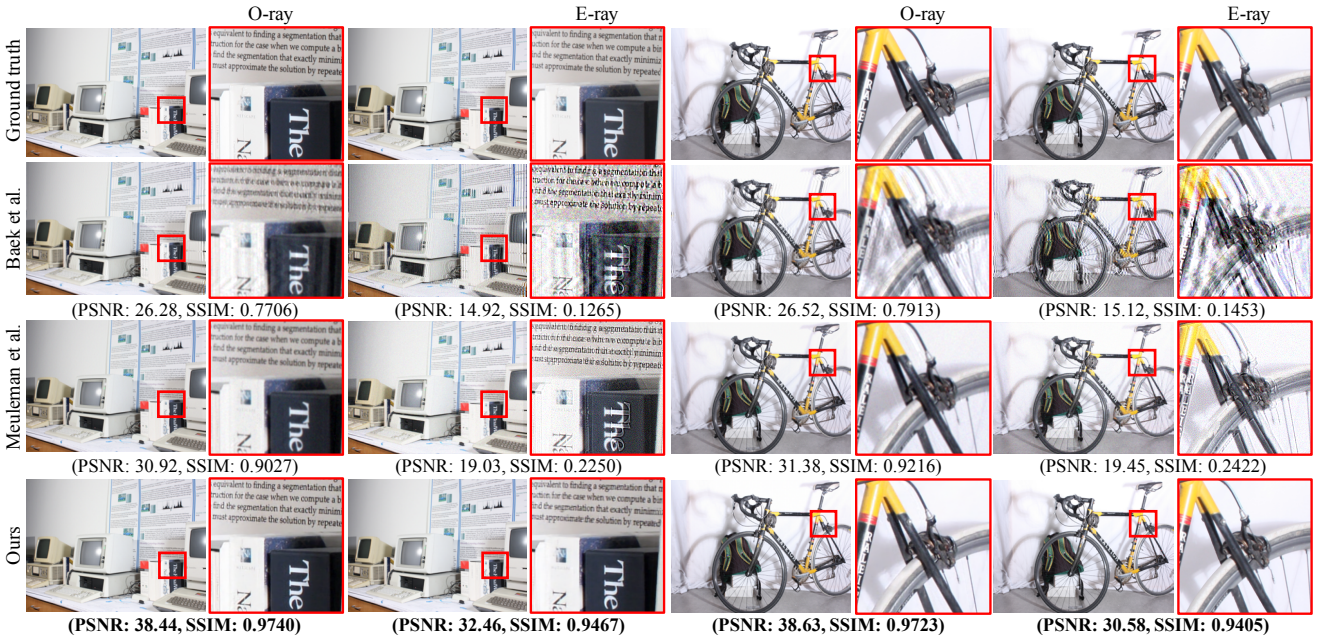


Figure 5: Comparison of our image restoration results with two birefractive stereo methods [1, 12] on synthetic double refraction images with the ground truth. Our method outperforms these two current methods.

	PSNR (dB)		SSIM	
	O-ray	E-ray	O-ray	E-ray
Baek et al. [1]	28.27	15.99	0.8062	0.1237
Meuleman et al. [12]	34.19	20.10	0.9158	0.2072
Our method	**39.80**	**34.51**	**0.9631**	**0.9379**

Table 1: Quantitative comparison of stereo image restoration quality with the synthetic test dataset. Bold texts mean the best accuracy.

restoration quality of o-ray and e-ray images. Furthermore, even our e-ray restoration results show a higher PSNR and SSIM values than the o-ray restoration results by the previous methods. Restored images are also visualized in Figure 5. It confirms that our method restores images with clearer edges and textures and removed noise effectively. In particular, our method obtains e-ray images without a typical deconvolution artifact of rings prevalent in the e-ray results restored by the previous methods.

Robustness. We evaluate the robustness of our network by running the restoration with image distortion by noise. We added Gaussian noise with different standard deviations on the images captured by our birefractive stereo camera and checked whether the stereo image is restored despite of the image degradation. The stereo restoration results of degraded input double refraction images for each noise level are shown in Figure 6. The highlighted image crops present that extracted rays from our network still preserves faithful geometrical displacements. This observation validates our methods' stability in stereo restoration to noise.

Also, we validate the robustness of our network trained with an attenuation ratio (τ=0.3) by evaluating reconstruction errors with image datasets simulated with different attenuation ratios $\tau \in \{0.15, 0.3, 0.45\}$. The average PSNR values of o-ray reconstruction results span in 36.02, 39.80, and 35.94 dB, respectively. The performance of our method is still higher than the previous works [1, 12] (Table 1).

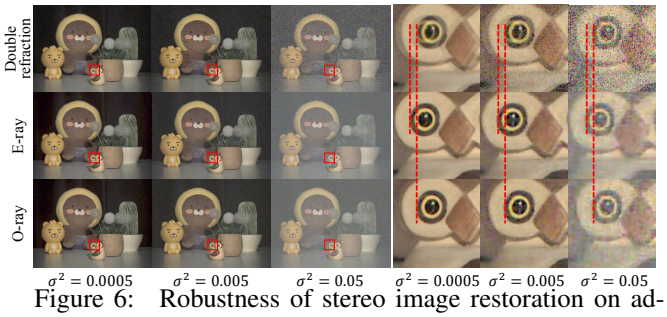


Figure 6: Robustness of stereo image restoration on additive Gaussian noise. Gaussian noise with variance of $\sigma^2 = 0.05, 0.005, 0.0005$ was added on the real scene.

4.1. Ablation Study

Two-Head vs. One-Headed Network. Our two-headed network architecture was compared with one-headed architecture to emphasize the importance of o- and e-ray joint estimation for stereo restoration with a balanced quality and appropriateness of the two-headed architecture of our stereo restoration network on the former purpose. We build a one-headed network that consists of one backbone network, one-headed network, and skip connection. The architecture of each component is equivalent to those of a two-headed network. One-headed network and two-headed network were trained in the same settings, equivalent to the first phase of

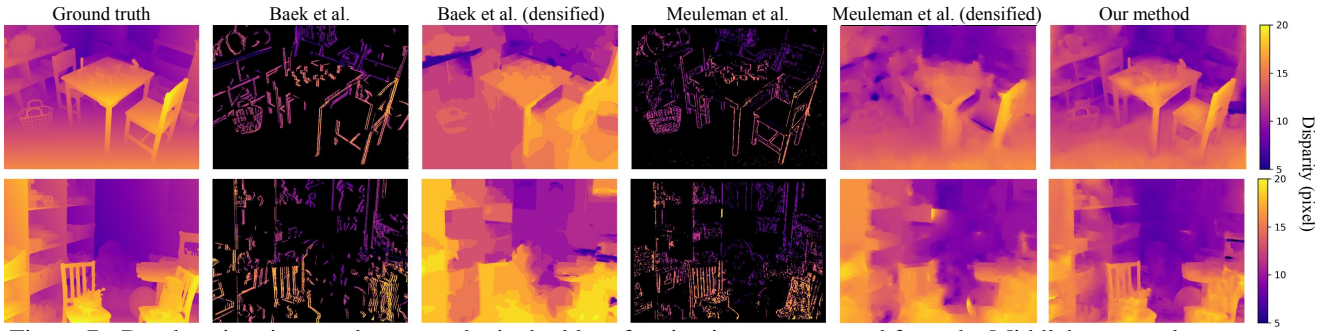


Figure 7: Depth estimation results on synthetic double refraction image generated from the Middlebury test dataset.

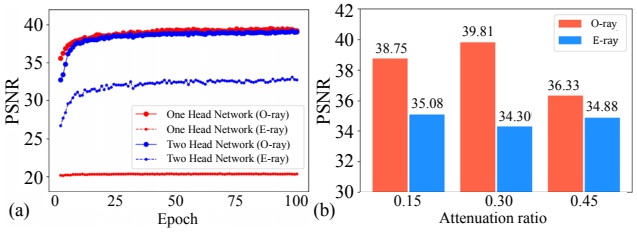


Figure 8: (a) Comparison on restoration performance of o-/e-ray between one-headed and two-headed networks. (b) Impact of the attenuation ratio on network training.

training described in implementation details (Section 3.2).

The difference is that the one-headed network was trained with o-ray supervision only, and an e-ray from the one-headed network was obtained by subtracting the output from a double-refraction image and rescaling it.

The performance of each network architecture for each epoch is illustrated in Figure 8(a). This plot shows that despite the stand-alone one-headed network restores o-ray in slightly better quality, simple subtraction of the restored o-ray from superimposed double refraction signals cannot restore the residual signals as compatible quality regardless of the o-ray restoration quality. Therefore, this experiment verifies that our network architecture design and training strategy of cooperatively restoring a pair of rays has successfully restored stereo images from double refraction.

Attenuation Ratio. We chose the attenuation ratio τ of our image formation model (refer to Equation (1)) upon the analysis of the former study [12] on the appropriate attenuation ratio for restoring both latent image and correct disparity at the same time. Yet, we further investigate the impact of the attenuation ratio on our network performance. We additionally synthesized double refraction data with different attenuation ratio ($\tau = 0.15, 0.45$). Supplemented ratios are consistent with the values investigated by Meuleman et al. [12].

We started from the pre-trained model obtained from the first phase of training and proceeded with the second phase of training with equal settings but with a newly generated dataset. Figure 8(b) portrays the PSNR of restored stereo images for each training stage of the corresponding attenuation ratios. The plot shows that the highest restoration quality of o-ray is obtained when $\tau = 0.3$ among examined values, which is consistent with the previous observation [12]. Still, it is notable that we obtained comparable o-ray restoration with a slighter gap of restoration quality between o-ray and e-ray when the attenuation ratio is 0.15.

5. Applications

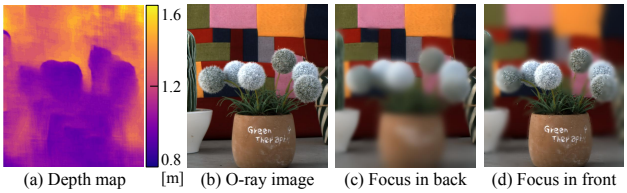
Dense RGB-D Imaging. We show that our stereo images restored from double refraction can be directly applied for dense depth map acquisition using existing stereo matching methods. We attached a PSM-Net [3] (pyramid stereo matching network) subsequent to our stereo restoration network as a depth acquisition module to estimate disparity between o-ray and e-ray. A double refraction image is fed into our stereo restoration network and then separated into o-ray and e-ray images. These outputs are fed into the PSM-Net after rectification. The mapping function for rectification was obtained by the previous calibration method [12].

We newly train the PSM-Net by feeding the o-ray and e-ray images acquired from the trained stereo restoration network and our synthesized double refraction dataset. The forward operation of the whole network pipeline ran while the parameters of our stereo restoration network were fixed. For the training phase, we used the l_1 smooth loss, which is consistent with the loss metric used for disparity regression by [3]. It is defined as:

$$\mathcal{L}_d = \begin{cases} 0.5\left(d_{gt} - \hat{d}\right)^2, & \text{if } \left|d_{gt} - \hat{d}\right| < 1 \\ \left|d_{gt} - \hat{d}\right| - 0.5, & \text{otherwise} \end{cases} . \quad (6)$$

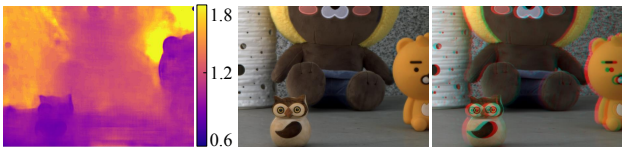
We used the Adam optimizer with parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$), and the initial learning rate is set to 10^{-3} for 8 epochs with the batch size of 32.

We validate the significance of disparity estimation with our restored stereo images by comparing disparity estimation by existing depth from double refraction methods [1, 12] on the synthetic test dataset. Since both methods produce a sparse depth map, we compared the disparity estimation error on the region that is valid in the previous methods and compared the estimation of dense maps by densifying the sparse depth map of [1, 12] with a popular diffusion algorithm [9].



(a) Depth map (b) O-ray image (c) Focus in back (d) Focus in front

Figure 9: Synthetic defocus. From the estimated depth map (a) and the reconstructed o-ray image (b), we render a defocused image, focusing on the background (c) and on the front object (d).



(a) Depth map (b) O-ray image (c) Anaglyph 3D photo

Figure 10: 3D Anaglyph. From the depth map (a) and the o-ray image (b), we generate an anaglyph image (c) with a baseline.

	Disparity RMSE (px.)		Bad pixel ratio	Sparsity
	Sparse	Dense	(%)	(%)
Baek et al. [1]	2.75	2.96	58.77	89.48
Meuleman et al. [12]	1.90	2.07	38.97	90.55
Our method	**1.44**	**1.71**	**19.79**	**0**

Table 2: Quantitative comparison of our dense depth estimation with other densified depth maps captured by other birefractive stereo methods. Bold texts mean the best value.

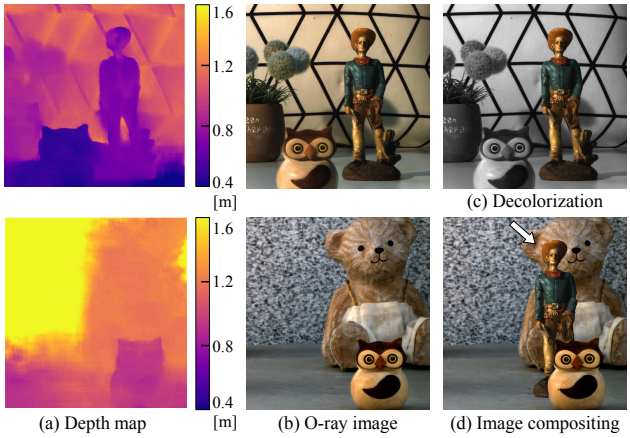
As shown in Figure 7, our method restores dense disparity map in great details. Moreover, disparity estimation errors and sparsity of depth estimations in Table 2 quantitatively verifies that disparity estimation with our restored stereo images gives more accurate results in sparse points and disparity map with better accuracy than densified disparity map from sparse values.

We also demonstrate qualitative depth estimation results of real scenes in Figures 9, 10 and 11. The actual depth values are computed using the disparity-to-depth converting equation: $d_{oe} = \frac{c}{z}$. The value of calibrated baseline using our prototype camera is $c = 6.756$ px·m. Refer to Section 2 in the supplemental document for the depth range.

Synthetic Defocus. Figure 9 shows a defocus application using our RGB-D results. From the dense depth map from the network and the RGB image of the o-ray, we synthesize a defocused image using the rendering technique proposed by Barron et al. [2]. As we already extracted a clean o-ray image and a depth map from a double refraction image, synthesizing a defocused image is straightforward.

Anaglyph Visualization. We generate anaglyph stereo images of red and blue filters in a target baseline using clean o-/e-ray images and a depth map, as shown in Figure 10.

Depth-based Object Segmentation. Our method enables depth-based object segmentation from a double-refraction image. Using a classical object segmentation algorithm, such as grab-cut [16], we extract objects from the captured scene. Figure 11 shows an example of segmentation used for background decolorization and depth-aware image composition. The occlusion regions are computed from the estimated depth without manual intervention.



(c) Decolorization

(a) Depth map (b) O-ray image (d) Image compositing

Figure 11: Depth-aware object segmentation. From the depth maps (a) and the o-ray images (b), we extract an object from the scene to render a decolorized image (c) and an occlusion-aware image composition (d).

6. Conclusion

We have proposed a *stereo restoration network* with a two-headed architecture and two different types of skip connections to effectively restore stereo images from double refraction using the geometrical relationship between two latent images overlapped in the double refraction image. Unlike previous image restoration and birefractive stereo studies, we restored both e- and o-ray, preserving the spatially variant displacements without relying on depth. Our results validate that our method restores both e- and o-ray with better quality in the PSNR metric. We further investigate the performance of our network on captured real scenes with a birefractive stereo camera and confirms that our network successfully generates stereo images from a monocular camera. Also, the robustness experiment with noise verifies the stability of our restoration method. Demonstrated applications, including dense RGB-D imaging, supports the practicality of our monocular stereo image restoration.

Acknowledgments

Min H. Kim acknowledges Samsung Research Funding Center of Samsung Electronics (SRFC-IT2001-04) for developing partial 3D imaging algorithms, in addition to a partial support of Korea NRF grants (2019R1A2C3007229), MSIT/IITP of Korea (2017-0-00072), and MSRA.

References

[1] Seung-Hwan Baek, Diego Gutierrez, and Min H Kim. Birefractive stereo imaging for single-shot depth acquisition. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016.

[2] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474, 2015.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[4] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Jianbo Jiao, Wei-Chih Tu, Shengfeng He, and Rynson WH Lau. Formresnet: Formatted residual learning for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–46, 2017.

[6] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5164–5173, 2020.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Ji✓rí Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[9] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM Transactions on Graphics (TOG)*, volume 26, page 70. ACM, 2007.

[10] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004.

[11] Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, and Mude Lin. Visually imbalanced stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Andreas Meuleman, Seung-Hwan Baek, Felix Heide, and Min H Kim. Single-shot monocular rgb-d imaging using uneven double refraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2465–2474, 2020.

[13] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[14] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[15] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[16] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[17] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Ne✓sić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.

[18] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] Takahiro Yano, Masao Shimizu, and Masatoshi Okutomi. Image restoration and disparity estimation from an uncalibrated multi-layered image. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 247–254. IEEE, 2010.

[20] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3555–3564, 2020.

[21] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2521–2529, 2018.

[22] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794, 2018.

[23] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[24] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.