This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CodedStereo: Learned Phase Masks for Large Depth-of-field Stereo

Shiyu Tan[1,*] Yicheng Wu[1,*] Shoou-I Yu[2] Ashok Veeraraghavan[1,†]
[1]Rice University [2]Facebook Reality Labs
{shytan, yicheng.wu, vashok}@rice.edu Shoou-I.Yu@fb.com

Abstract

Conventional stereo suffers from a fundamental trade-off between imaging volume and signal-to-noise ratio (SNR) – due to the conflicting impact of aperture size on both these variables. Inspired by the extended depth of field cameras, we propose a novel end-to-end learning-based technique to overcome this limitation, by introducing a phase mask at the aperture plane of the cameras in a stereo imaging system. The phase mask creates a depth-dependent yet numerically invertible point spread function, allowing us to recover sharp image texture and stereo correspondence over a significantly extended depth of field (EDOF) than conventional stereo. The phase mask pattern, the EDOF image reconstruction, and the stereo disparity estimation are all trained together using an end-to-end learned deep neural network. We perform theoretical analysis and characterization of the proposed approach and show a $6\times$ *increase in volume that can be imaged in simulation. We also build an experimental prototype and validate the approach using real-world results acquired using this prototype system.*

1. Introduction

Stereo-based 3D reconstruction, while extremely popular, suffers from a fundamental trade-off between volume of imaging and noise. If you want to retain a large volume of imaging, then in order to get sharp texture features for correspondence, you need to ensure that the depth of field (DOF) of the cameras covers the entire volume. This necessitates a narrow aperture, rapidly reducing the total light level reaching the sensor (since SNR is quadratically related to aperture size). As a consequence, it is challenging to get large volume, high quality, and high resolution stereo-based 3D reconstruction in light-limited environments.

In light-limited environments, typically either the exposure duration or the aperture size of a camera is increased to increase light throughput. But for scenarios where there is either scene motion (eg., motion capture) or camera motion (eg., robotics, autonomous navigation), increasing exposure duration results in motion blur. On the other hand, increasing the aperture size will result in a smaller depth of field, thereby reducing the volume that can be reconstructed.

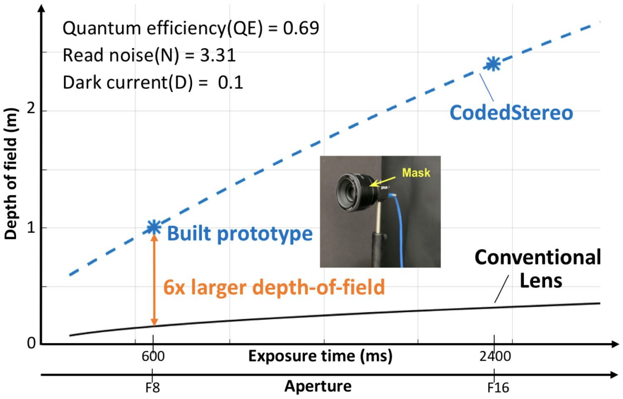


Figure 1. **Tradeoff between depth of field and aperture size on simulated data.** We propose a CodedStereo system that can provide an $6\times$ increase in DOF (blue dashed curve, with stars for particular observations). In the curves, the x-axis is linearly sampled in exposure time, and the corresponding f-numbers are converted to maintain the same SNR level of 50dB. Both our system and the conventional lens are focused at $1m$, with a $50mm$ focal length.

Inspired by the extended depth of field (EDOF) imaging techniques [10], we present CodedStereo, a technique in which we add optimized phase masks to the aperture of each of the two stereo cameras. These phase masks allow each camera to maintain a large aperture size, increasing the light throughput of the cameras. Meanwhile, the phase masks are specially designed to produce a depth-dependent focal blur that allows back-end stereo algorithms to continue to retain high resolution and quality. In addition, our learned phase mask not only enables more accurate depth estimation, but also encourages sharper extended depth of field RGB images that can be used for downstream applications such as point cloud colorization. The reconstruction algorithms and the phase masks are simultaneously optimized using an end-to-end learning framework. The main technical contributions of this paper are:

I. We propose CodedStereo, a technique to recover large-

[*] These two authors contributed equally. [†] Corresponding author.

volume, high-quality, and high-resolution 3D reconstructions in light-limited environments. The key idea in CodedStereo is the introduction of a phase mask in the aperture of the stereo cameras that allows us to increase the aperture size of the cameras without sacrificing the depth of field.

II. We develop an end-to-end learning framework to jointly optimize the phase masks and the algorithms both for RGB image and disparity reconstructions.

III. We demonstrate the significant performance benefits of CodedStereo both in simulation and using a prototype system.

2. Related work

Stereo matching. Given two (or more) cameras looking at the scene from different perspectives, the goal of stereo is to find the corresponding scene points between the two camera views and use this to estimate depth based on triangulation. Traditional methods[29, 16] typically formulate it as a multistage optimization problem, including matching cost computation, cost aggregation, disparity optimization, and post-processing. Recently, learning-based stereo algorithms have become popular primarily due to their improved performance. Many networks, inspired by the traditional stereo matching pipelines, have been shown to achieve state-of-art results [27, 4, 23, 11, 38]. Among these algorithms, [23, 11] are computationally efficient and can be used for real-time inference. However, it is well known that existing stereo algorithms degrade in performance when the images contain significant blur or noise [22, 18].

Low light stereo. Extending stereo algorithms and improving their performance in the presence of significant noise (as is the case in low-light imaging) is an area of active research. The simplest solutions attempt to first denoise the stereo pairs before the correspondence search. But unlike generic denoising algorithms [5, 21], these methods pay more attention to the consistency of the denoised image pairs to make sure the stereo matching algorithms can find the corresponding features. Another technique to improve low-light performance is to replace one or both of the stereo cameras with monochrome sensors, resulting in approximately a $3\times$ increase in light throughput [18].

Stereo and defocus blur. When the camera aperture is large such that the scene is no longer contained within the depth of field of the camera, focus blur is apparent in the captured images. There have been attempts to exploit this focus blur as an additional depth cue to compensate for the degraded stereo performance [6, 34, 7, 13]. Furthermore, Takeda *et al*. [33] proposed the addition of amplitude masks in the aperture plane. The use of amplitude masks increases the variations in the depth-dependent blur, improving depth from defocus approaches. Our technique also proposes the addition of a mask within the camera's aperture plane. However, there are two key differences. First, since our main goal is low-light imaging, we use phase masks instead of amplitude masks to obtain high light throughput. Second, compared to the heuristic mask design in [33], the proposed design is directly optimized based on the 3D reconstruction, which improves the performance.

Extended depth of field imaging. For a conventional camera, it is well understood that the aperture size controls the relationship between the depth of field and SNR. Larger apertures result in higher light collection leading to an increase in SNR, but at the cost of decreasing the depth of field. There have been a host of techniques that have been developed to maintain a large aperture and a large depth-of-field. One idea that has emerged from this line of inquiry is to reconstruct all-in-focus images from integrated images with a shaking sensor [25]. Another key idea is the use of a phase mask in the aperture plane to control the depth-dependent blur in a manner that makes the resultant blur invertible [10, 8, 9, 12, 31]. Our design is intimately related to these efforts and the main difference is that when applying these techniques to stereo, one must pay attention to maintaining consistency across views, so that correspondence matching algorithms remain stable.

End-to-end mask design. Over the last few years, several techniques have emerged where optical system design parameters and reconstruction algorithms are jointly optimized in an end-to-end manner. The primary rationale for this end-to-end learning framework is the significant improvements that are obtained as a result of this joint optimization. Such methods have been shown to achieve superior performance in demosaicing [2], monocular depth estimation [35, 3, 15], microscopy [26, 19], structured light [1, 36], EDOF [31], and high dynamic range [24, 32] imaging. Our technique is of a similar vein, but tackling the problem of large volume, low-light stereo reconstruction.

3. Imaging Volume vs SNR: The tradeoff

Traditional stereo exhibits a fundamental trade-off between light level, exposure time, and volume of reconstruction that limits the quality of 3D reconstructions. In a traditional camera, the image signal to noise ratio (SNR) is proportional to incident light intensity, which in turn is proportional to the product of aperture area, illumination level, and exposure duration. Thus,

$$\mathrm{SNR} \propto \frac{L_s T D^2}{\sigma_{tot}} = \frac{L_s T f^2}{\sigma_{tot} F_{\#}^2}. \tag{1}$$

where $D = f/F_{\#}$ is the aperture diameter, f is the focal length, $F_{\#}$ is the f-number, T is the exposure duration, L_s denotes the average light intensity, and σ_{tot} refers to the total noise. This relationship indicates that the imaging SNR will become extremely low either under low-light conditions or when scene/camera dynamics require the use of short exposure durations.

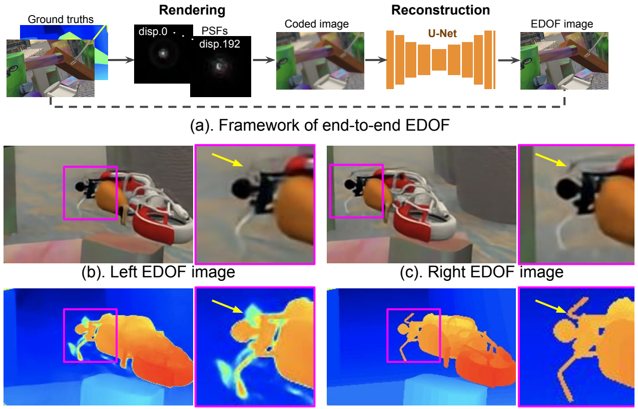
Imaging SNR, in turn, impacts the quality of correspondence between the stereo pair, resulting in low-quality and/or low-resolution 3D reconstructions. The simplest solution to improve imaging SNR is to use a larger aperture. Unfortunately, this is typically not a feasible solution for stereo systems since this reduces the depth of field, which, as a result, significantly reduces the imaging volume for accurate 3D reconstruction. Specifically, the approximate depth of field (DOF) can be determined by focal length f, distance to subject z_0, acceptable circle of confusion size c, and aperture diameter [17].

$$DOF \propto \frac{2z_0^2 F_{\#} c}{f^2}. \tag{2}$$

Figure 1 demonstrates this tradeoff between depth of field and aperture size (or F#) for a lens with a focal length of $50mm$ and focused $1m$ in front of the lens. This shows that if you want a large imaging volume, then you need to use a narrow aperture. In particular, achieving a DOF of $1m$ requires the use of a $F32$ aperture. Unfortunately, under low-light conditions and/or with short exposure duration, such a small aperture size would severely limit light throughput resulting in extremely noisy images – which in turn will result in low-quality 3D reconstructions. On the other hand, a large aperture in search of better light throughput induces significant focus blur within the imaging volume. This blur again affects the quality of stereo correspondences and 3D reconstruction. Inspired by the extended depth of field imaging techniques, the key idea in our approach is to utilize a phase mask at the aperture plane to 1) keep the aperture large, and 2) create depth-dependent yet numerically invertible focal blur point spread functions that allow for high-quality 3D reconstruction over the entire imaging volume. The increase in imaging volume/DOF achieved by CodedStereo is shown for comparison in Figure 1.

4. Extended depth of field in stereo matching

One naïve technique to overcome the imaging volume vs SNR tradeoff in stereo systems would be to replace each of the cameras in a stereo system with an EDOF camera. Surprisingly, this naïve application of EDOF does not seem to result in significant improvement to the imaging volume in stereo systems. The primary reason for this discrepancy is that the deconvolution algorithms, irrespective of whether they are optimization-based [10, 8, 9, 25] or learning-based [31, 12], produce minor inconsistencies across views. While these inconsistencies are imperceptible and do not seem to affect the perceptual quality of individual images, they have a significant effect on the stereo correspondence search. As a result, the stereo correspondence search produces significant errors, affecting the quality of the 3D reconstructions.



(a). Framework of end-to-end EDOF

(b). Left EDOF image

(c). Right EDOF image

(d). Prediction from EDOF pairs

(e). Ground truth

Figure 2. **Naïve EDOF stereo results in 3D reconstruction artifacts due to feature inconsistencies.** (a) The framework used to learn the e2eEDOF phase mask. (b)-(c) Reconstructed EDOF images with inconsistent fine features across views. (d) Predicted disparity map from EDOF pairs. The algorithm failed to recover stereo correspondence for unmatched regions. (e) Ground truth.

Figure 2 shows the effect of these imperceptible inconsistencies on stereo reconstructions. We follow the technique in [31] to learn an optimal phase mask for EDOF imaging and use that same phase mask for both the left and the right camera in a stereo system. The e2eEDOF learning framework and the prediction results from EDOF pairs are shown in Figure 2. A close inspection of the results shows that the matching algorithm failed to recover correspondence due to the inconsistencies in the individual EDOF recovered images, as pointed to by the yellow arrow.

5. CodedStereo framework

Our technique consists of a single optimized phase mask inserted into the aperture of both the cameras in a stereo pair. With these phase masks inserted, the aperture of these cameras can remain wide open, allowing significantly larger light collection thereby improving imaging SNR. The depth-dependent blur caused by the insertion of these phase masks is jointly optimized along with the disparity and image reconstruction algorithms to maximize the volume-SNR tradeoff in stereo. We call our technique 'CodedStereo'. Our system simultaneously obtains sharp image texture and stereo correspondence in a large depth of field, without sacrificing SNR or light throughput.

As shown in 3, the end-to-end training pipeline consists of three distinct parts: (a) *Rendering:* A RGB-Disp simulator to render left/right coded images using texture and depth as inputs (while accounting for the depth-dependent defocus effect of a particular phase mask), (b) *Disparity Prediction:* a DispNet-based deep network to estimate disparity from coded pairs, and (c) *RGB Image Reconstruction:* a U-Net to reconstruct sharp images. The detailed description of

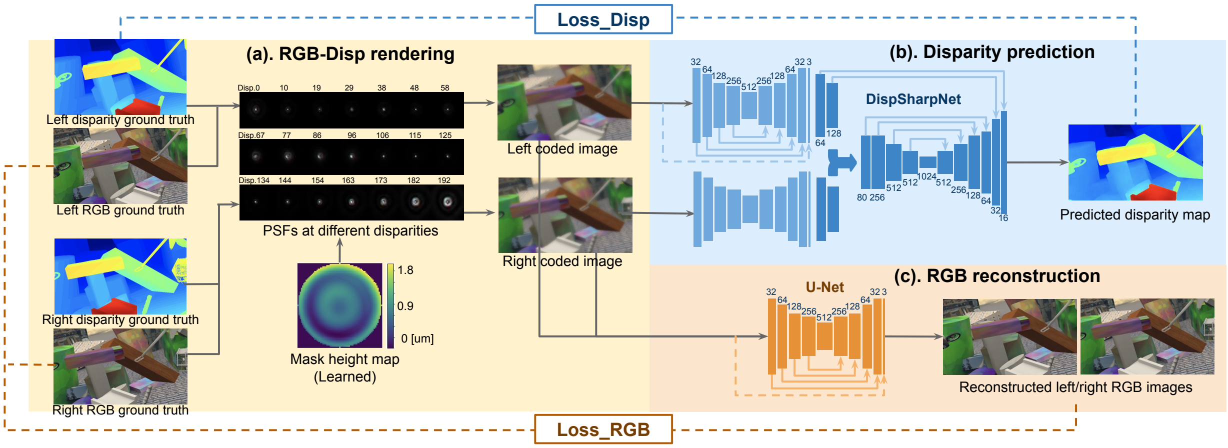


Figure 3. **Framework overview**. We learn the phase mask together with a disparity prediction network and an RGB reconstruction network in an end-to-end manner. In the RGB-Disp rendering layer, disparity-dependent PSFs are first simulated given the learnable phase mask. These PSFs are then convolved with ground truths to render left/right coded images, which are the inputs to the following reconstruction networks. We use a DispNet-based network and a U-Net-based network to estimate the sharp texture images and the disparity map, respectively. The loss of reconstructed texture and disparity are summed up together in backpropagation to update the mask height map and the network parameters at the same time.

each component is discussed in the following subsections.

5.1. Rendering Using RGB-Disp Simulator

In conventional stereo, it is assumed that the entire scene is within the depth of field. When this is not true, as is the case here, the defocus blur apparent on the captured images depends upon the depth of the scene point, and thus depends upon the disparity between the corresponding points of two camera views. In addition, when a phase mask is inserted into the aperture plane, the disparity-dependent point spread function (PSF) also depends upon the phase mask pattern. The goal within the rendering layer is to accurately model the effect of phase mask pattern and disparity on the captured left and right images in a CodedStereo system.

The RGB-Disp rendering is based on Fourier optics theory [14] and is fully differentiable to enable end-to-end training. We first simulate the point spread functions (PSFs) for each camera as the squared magnitude of the Fourier transform of the pupil function (which depends on the phase mask pattern)

$$PSF_\lambda \propto |\mathcal{F}\{A\exp(\phi^M + \phi^{DF})\}|^2, \tag{3}$$

where λ is the wavelength. In the pupil function, A denotes a circular amplitude function with respect to aperture size, ϕ^M denotes the phase shift induced by the phase mask (proportional to the mask height map), and ϕ^{DF} denotes the defocus phase. The defocus phase can be further derived as a function of disparity d,

$$\phi^{DF}(x_1, y_1) = \frac{k_\lambda}{2fb}(d - d_0)(x_1^2 + y_1^2) \tag{4}$$

where $k_\lambda = 2\pi/\lambda$ is the wavenumber, f is the focal length, and b is the baseline between two views. (x_1, y_1) denotes the spatial coordinate on mask plane, and d_0 is the corresponding disparity value at in-focus depth. We then render the coded images by convolving the ground truth RGB textures with the disparity and wavelength-dependent PSFs.

$$I_\lambda^c = \sum_d M_d \cdot (I_\lambda^i * PSF_{\lambda,d}) + noise \tag{5}$$

where $\cdot$ is an element-wise product operator, I^i is the input all-in-focus image, and I^c is the rendered coded image. M_d denotes a segmentation mask (1 when the pixel disparity is d, 0 otherwise). To account for boundary occlusions, the segmented layers were further blended using the normalized matting weights[20]. To render the effect of noise (which would be significant under low-light conditions), we apply an additive Gaussian noise, whose standard deviation is calculated based on the aperture size, light-level, and exposure duration.

5.2. RGB and Disparity Reconstruction

We use two separate networks to reconstruct the disparity map and sharp texture images. The texture reconstruction network is based on a modified residual U-Net [28] in which the differences between coded image and ground truth image (i.e. residual image) are learned. The advantage of learning a residual image is to encourage high-frequency information recovery, like edges and detailed textures, and therefore such residual learning techniques are widely used in per-pixel estimation problems such as deblurring.

For disparity prediction, we adopt the structure of DispNetC [23]. Note that DispNetC only outputs disparity maps at half the resolution of the input stereo pairs. We modify

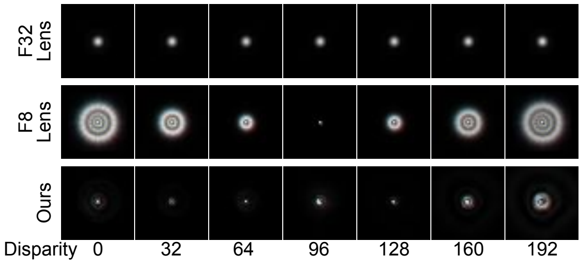


Figure 4. **Optimized PSFs in simulation (focus at 96px).** Each PSF slice is normalized for better visualization. PSFs for conventional $F32$ and $F8$ lenses are also shown for comparison.

it by adding extra deconvolution layers to upsample the disparity map [27], so that the final output is the same size as the input left/right coded images. We further found that an extra encoder-decoder module before DispNetC can benefit the feature extraction of stereo pairs especially in areas of the image with a large amount of out-of-focus blur. Therefore, we process coded left/right images separately through a shared-weighted encoder-decoder layer followed by two convolution layers to extract features, and then horizontally correlate the features. We considered a maximum shifting of 64 pixels which corresponds to 192 pixels in the original coded images. We call our disparity prediction network DispSharpNet, as it enables disparity estimations with extra details and sharper boundaries. More details of network architectures are shown in the supplementary.

5.3. Implementation details

We optimized the phase mask over a depth range of $[0.7m-1.7m]$ for a stereo system with a baseline of $22mm$. The lenses are focused at $1m$ with focal lengths of $50mm$ and $F8$ aperture sizes. The sensors' pixel size was set to $4.8\mu m$, corresponding to a disparity range of $[134-326]$ pixels. To avoid large disparity values, we manually pre-shifted the right image by 134 pixels to the right. This is equivalent to reduce the disparities by 134, and thus the disparity range changes to $[0-192]$. During training, the mask was directly optimized over the reduced disparities (21 distinct values sampled in $[0-192]$). The learnable mask height map was discretized with a pitch size of $88\mu m$ at a resolution of 71×71. Similar to the previous works [30, 35], we further parameterized the height profile and represented it using Zernike polynomials with 55 coefficients.

Loss function. During training, the loss function is defined as a combination of disparity prediction error and RGB reconstruction error. We made use of the root mean squared error for both the estimated RGB image $\hat{I}$ and the predicted disparity $\hat{d}_i$ at different resolutions i.

$$\begin{aligned} Loss &= Loss\_Disp + Loss\_RGB \\ &= \frac{1}{\sqrt{M}}\sum_i \alpha_i \left\|d_i-\hat{d}_i\right\|_2 + \frac{1}{\sqrt{N}}\gamma(\left\|I^l-\hat{I}^l\right\|_2 + \left\|I^r-\hat{I}^r\right\|_2), \end{aligned} \tag{6}$$

where α_i, γ are the corresponding weights, and M, N are the number of pixels in the RGB image and disparity map, respectively. l denotes the left, and r denotes the right. For stable feature matching, similar to DispNet [23], we adopted a loss weight schedule to start training with only the lowest resolution loss, and progressively increase the weights of losses with higher resolutions.

Dataset & training. Our model was end-to-end trained on a synthetic dataset consisting of dense ground truth disparity maps (enabling our RGB-Disp rendering) for 35,454 training and 4370 testing stereo pairs [23]. During training, the image patches were randomly cropped into a size of 384×768, and preprocessed by subtracting out their means and dividing by their standard deviations. We optimized our phase mask and network parameters using Adam optimizer ($\beta_1=0.9$, $\beta_2=0.999$) with a batch size of 8 for 50 epochs, on GeForce RTX 2080 Ti GPUs.

6. Simulation results

We conducted quantitative and qualitative evaluations of our method in simulation. The phase masked learned with γ=0.5 was selected for evaluations, both in simulation and in experiment, as it simultaneously produces the sharp RGB texture and accurate disparity map over a large depth of field. The optimized PSFs are shown in Figure 4. Compared to a conventional $F8$ lens, our PSFs have a significantly shrunken radius of the Airy disk (comparable or even smaller than $F32$ lens) at out-of-focus depths, improving the reconstruction of both RGB images and disparity map with high resolution. Furthermore, our PSFs also come with some variations along the disparity axis, providing complementary blur cues to assist the disparity prediction of problematic areas.

Comparison with conventional lenses. To illustrate the improvement of our system over conventional designs, we compared our masks with a pair of $F32$ conventional lenses (small-aperture resulting in low SNR), and a pair of $F8$ lenses (open-aperture with a large amount of out-of-focus blur). For each system, the networks were trained with an additive 2% Gaussian noise, assuming the cameras are all designed to work under normal-light conditions.

The average peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) are adopted for evaluations on the texture reconstruction, and the end-point error (EPE) and the 3-pixel threshold error rate (3px) are used for the disparity, as shown in Table 6. Our method outperforms conventional designs with higher RGB reconstruction accuracy and lower disparity prediction error. A visual comparison is shown in Figure 5. It is clear to see that the $F32$ system suffers from low SNR, resulting in noisy textures and disparity maps, while the $F8$ system fails to reconstruct fine features due to out-of-focus blur. Our design outperforms the $F32$ system and the $F8$ system in a high-quality,

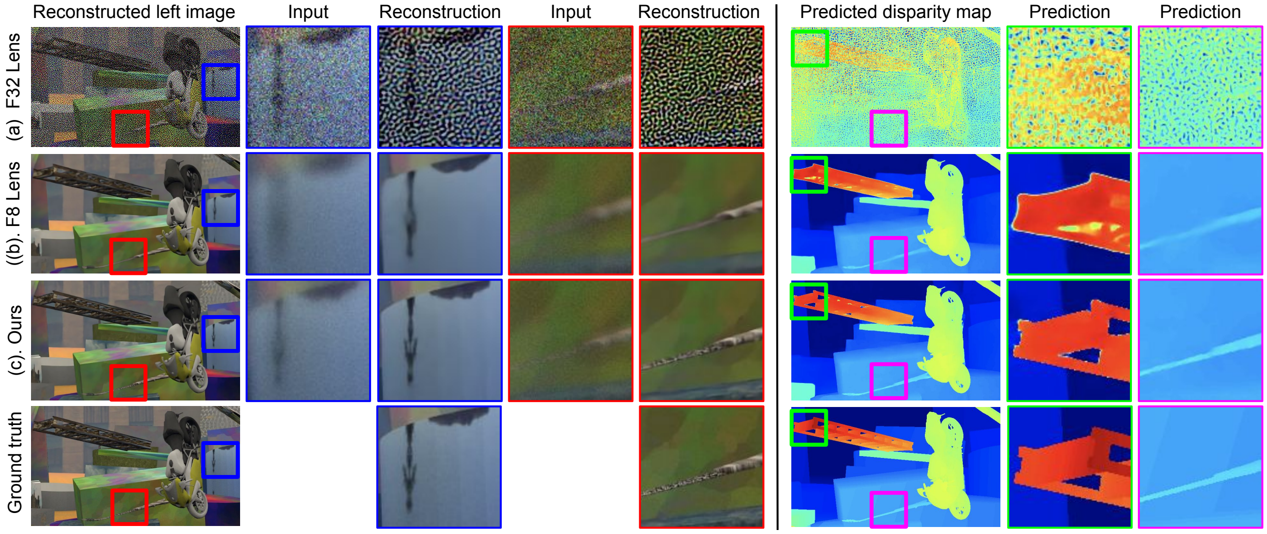
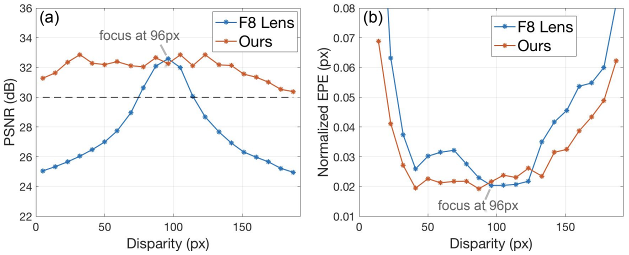


Figure 5. **Comparison with conventional baselines (in simulation)**. (a) with small-aperture conventional $F32$ lenses. (b) with open-aperture conventional $F8$ lenses. (c) ours with optimized masks. For comparison, we applied the same reconstruction networks to $F32$ and $F8$ systems as ours, i.e. U-Net for RGB images estimation, and DispSharpNet for disparity prediction. Results show that our design outperforms conventional designs in a high-quality, high-resolution reconstruction with clear details and sharp edges.



		F32 Lens	F8 Lens	Ours
RGB	PSNR[dB]	11.27	28.52	**31.90**
	SSIM	0.048	0.807	**0.880**
Disp.	EPE[px]	38.034	1.815	**1.512**
	3px[%]	95.45%	9.79%	**7.85%**

Figure 6. **Comparison with conventional lenses**. Top: The reconstructed PSNR and EPE (normalized to disparity ground truth) variations with disparity are plotted. Our method is significantly better than conventional baselines, especially at the out-of-focus range, resulting in a 6× increase in depth of field (black dashed line for PSNR threshold at 30dB). Bottom: Average PSNR and SSIM are used for evaluations on texture reconstruction (the higher the better), and average EPE and 3-pixel error rate are used for evaluations on disparity prediction (the lower the better).

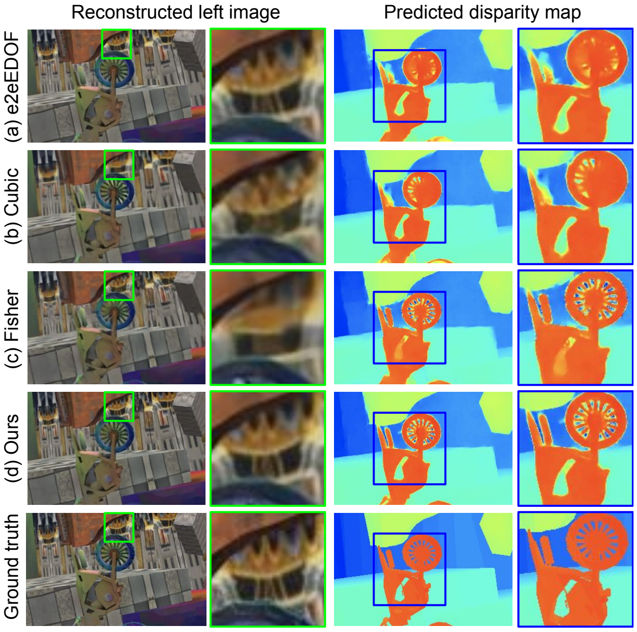
high-resolution reconstruction with clear details and sharp edges. We further compared the depth of field of the $F8$ system and ours, by analyzing the reconstruction PSNRs over disparities, as shown in Figure 6. Our methods surpass the PSNR threshold (30dB) for all the disparities within the range [0-192], resulting in a 6× increase in depth of field (invert disparity) compared to the $F8$ system. The curves of the normalized disparity EPE (EPE divided by the ground truth) are also shown on the right, indicating our disparity prediction improvement in the out-of-focus range.

Comparison with other masks. We further compared our method with several other coded-aperture stereo systems. These coded masks were optimized based on the theoretical or heuristic properties of the PSFs. Specifically, the Fisher mask was designed to increase the PSFs variation over depth using Fisher information [30], while the cubic mask was derived to force the PSFs to be similar over a large depth range [10]. The comparison results are shown in Figure 7. Reconstruction results of the e2eEDOF mask (Sec. 4) are also shown in the figure. Our optimized mask outperforms the e2eEDOF mask, the Fisher mask, and the cubic mask for both RGB and depth estimation.

Ablation study. As mentioned in Sec. 5.3, the overall loss contains both the loss of RGB reconstruction and the loss of depth prediction, and γ is the corresponding weight. In Table 1, we compared the performance under different γ values. As expected, the network performs good depth estimation when γ is small, and on the contrary, when γ is large the network performs good RGB estimation. We finally chose $\gamma = 0.5$ in our system.

7. Real experiment

To demonstrate our method, we built a hardware prototype with a fabricated mask inserted in a Yongnuo $50mm$ lens (with a $F8$ aperture). As shown in Figure 8, a Blackfly (BFS-U3-200S6C-C) color camera with $2.4\mu m$ pixel size was used as the sensor. To match simulations, we sub-



		e2eEDOF	Fisher	Cubic	Ours
RGB	PSNR[dB]	**32.44**	27.85	29.84	31.90
	SSIM	**0.880**	0.820	0.839	**0.880**
Disp.	EPE[px]	2.051	1.834	1.929	**1.512**
	3px[%]	11.71%	10.43%	10.28%	**7.85%**

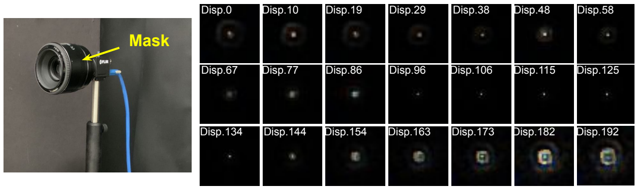
Figure 7. **Comparison with other masks in simulation**. The e2eEDOF mask is end-to-end trained [31], and its disparity is directly estimated from EDOF image pairs. The disparities of Fisher [30] and Cubic [10] masks are predicted from coded images. Our CodedStereo mask outperforms others on disparity estimation, and has comparable texture reconstruction accuracy to EDOF.

sampled the sensor pixels by 2 × 2 so that the equivalent pixel size is $4.8\mu m$ (with a resolution of 1824×2736). The left/right coded image pairs were captured by translating the camera $22mm$ (baseline) using a Thorlabs linear stage. Similar to simulation settings, scenes were constructed within a volume of $[0.7m - 1.7m]$ from the prototype and the captured right images were pre-shifted by 134 pixels to reduce the disparity value. The reduced disparity range then drops to $[0 - 192]$, aligning with the settings for which the network was trained.

Mask fabrication & system calibration. We fabricated our mask using two-photon lithography (Photonic Professional GT Nanoscribe 3D printer). During printing, the height-map of the mask was discretized (in height) into 10 steps with a stepsize of $200nm$. To account for any imperfection and misalignment in real experiments, we calibrated the PSFs with a deconvolution-based algorithm inspired by [37, 35]. The calibrated PSFs are shown in Fig. 8, which are used to finetune the reconstruction networks for best performance. More fabrication and calibration details can be found in the supplemental material.

		γ=0	γ=0.25	γ=0.5	γ=∞
RGB	PSNR[dB]	28.82	30.34	31.90	**32.44**
	SSIM	0.842	0.874	**0.880**	**0.880**
Disp.	EPE[px]	**1.462**	1.477	1.512	1.718
	3px[%]	7.73%	**7.25%**	7.85%	9.13%

Table 1. **Ablation study on various γ values in loss function.** PSNR, SSIM of RGB reconstruction and EpE, 3-pixel error rate of disparity prediction as a function of γ.



(a). Our prototype

(b). Calibrated PSFs of our prototype

Figure 8. **Built prototype with calibrated PSFs.** We fabricated the mask and built a prototype to demonstrate our method.

Experiment results. Our real-world experiment results are shown in Fig. 9. From the captured coded image pair, our method can reconstruct both RGB image and disparity with high accuracy in a large depth-of-range. Similar to the simulation section, we further compared our prototype with conventional $F8$ and $F32$ lenses in real experiments. The same exposure time (600ms) was applied for all three settings. As a reference, we included the reconstruction results of a $F32$ system with 10s exposure time to show the best result we can get without the SNR constraint. The results are shown in Fig. 10. As expected, the $F32$ system produces noisy reconstructions given low SNR, while the $F8$ system fails to recover fine features in texture and disparity due to the large out-of-focus blur. Our CodedStereo system generates high-quality results similar to the long-exposure $F32$ system with significantly shorter exposure time.

8. Conclusion & discussion

In this paper, we proposed a CodedStereo system that can recover large-volume, high-resolution 3D information under light-limited environments. The key idea of our system is to introduce a single phase mask at the aperture plane of stereo cameras. The mask was end-to-end learned together with an RGB reconstruction network and a disparity estimation network. The optimized phase mask creates a disparity-dependent point spread function, allowing us to recover sharp image and stereo correspondence over a significantly expanded depth of field than conventional stereo. We showed in simulation and experiments (with a prototype) that our method outperforms conventional lens and heuristic masks on both reconstructed texture and disparity.

Despite the advantages of our method, some limitations remain. First, the introduction of the phase mask makes the hardware system more complicated in design, and the

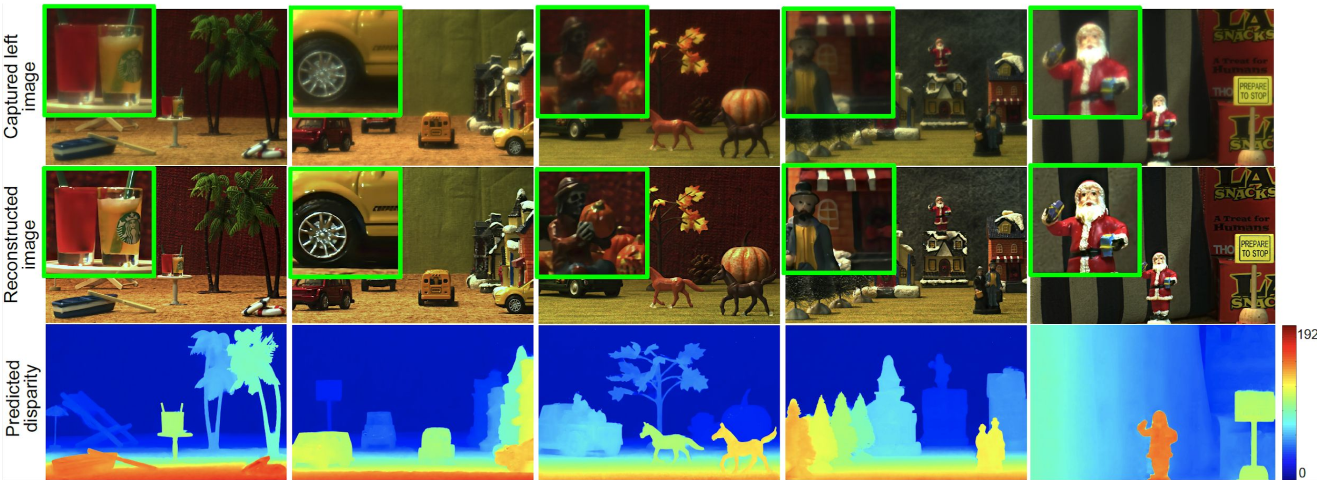


Figure 9. **Experiment results of various real-word scenes using our CodedStereo prototype.** Reconstruction results are shown for real scenes with both uniform background and non-uniform background (the last column, variation in texture/depth).

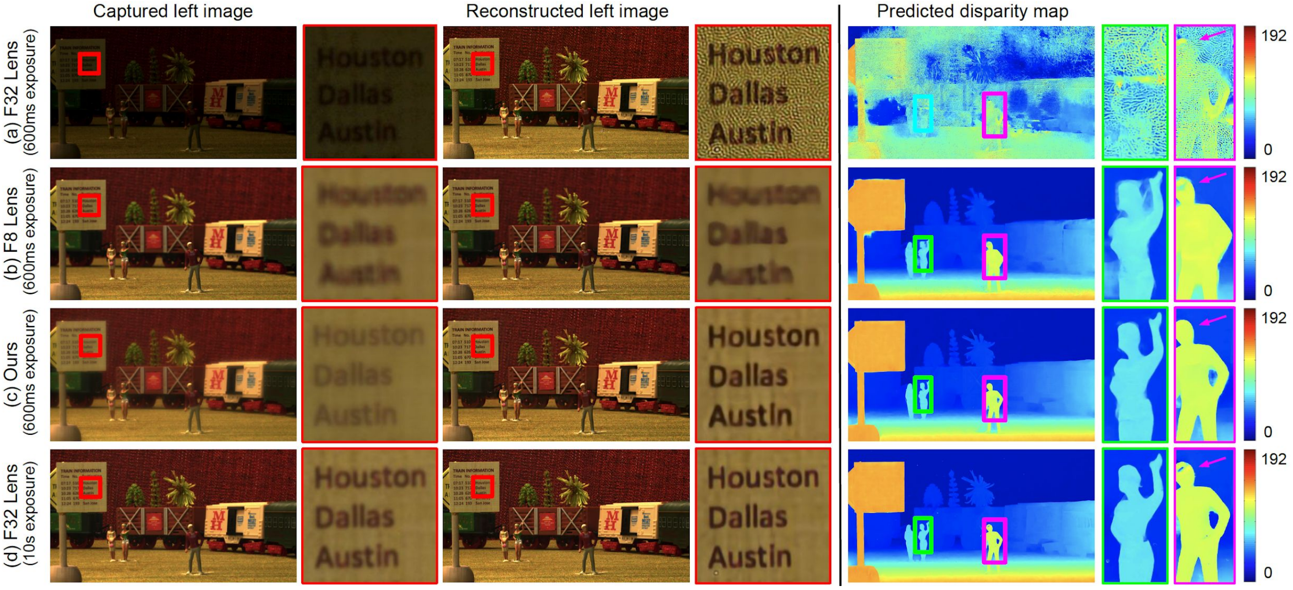

Captured left image
Reconstructed left image
Predicted disparity map
(a) F32 Lens (600ms exposure)
(b) F8 Lens (600ms exposure)
(c) Ours (600ms exposure)
(d) F32 Lens (10s exposure)
Houston
Dallas
Austin
192
0

Figure 10. **Comparison with conventional lenses in real-world experiments**. We compare the real-world performance of our prototype to the traditional $F32$ and $F8$ lenses here. The coded images of (a)-(c) are captured in the same 600ms exposure (scaled up by 8 times for $F32$ for visualization). (d) is long-exposure (10s) captured with a $F32$ lens, and the reconstructions are considered as the ground truths. As predicted by simulation, our system is superior to conventional designs in RGB and disparity reconstruction, and outperforms all these baselines (even long-exposure $F32$) on the disparity prediction in saturated regions, as pointed by the pink arrow.

re-training of phase masks and networks are required for different system settings (such as the lens focal length, the aperture size, the focus depth or the sensor pixel size that ends up with different defocus blur/disparities). Second, since our method is based on depth from disparity/defocus methods, it inherits their limitations on texture-less areas. Moreover, there is a trade-off between the accuracy of disparity and texture reconstruction (controlled by the weight γ). Further optimizing the system design might can mitigate this trade-off, including designs with two different phase masks/lenses across two views. Looking into the future, we hope to extend our framework to multi-view large depth-of-field stereo, enabling more reliable 3D information capturing under low-light conditions.

Acknowledgement. We would like to thank Hernan Badino and Matt Stewart at Facebook Reality Labs for their helpful discussions. AV, ST and YW were partially supported through a research grant from Facebook Reality Labs and the NSF CAREER Award #1652633.

References

[1] Seung-Hwan Baek and Felix Heide. Polka lines: Learning structured illumination and reconstruction for active stereo. *arXiv preprint*, 2020. 2

[2] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. In *NeurIPS*, 2016. 2

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *ICCV*, 2019. 2

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 2

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 2

[6] Ching-Hui Chen, Hui Zhou, and Timo Ahonen. Blur-aware disparity estimation from defocus stereo images. In *ICCV*, 2015. 2

[7] Zhang Chen, Xinqing Guo, Siyuan Li, Xuan Cao, and Jingyi Yu. A learning-based framework for hybrid depth-from-defocus and stereo matching. *arXiv preprint*, 2017. 2

[8] Oliver Cossairt and Shree Nayar. Spectral focal sweep: Extended depth of field from chromatic aberrations. In *ICCP*, 2010. 2, 3

[9] Oliver Cossairt, Changyin Zhou, and Shree Nayar. Diffusion coded photography for extended depth of field. In *SIGGRAPH*. 2010. 2, 3

[10] Edward R Dowski and W Thomas Cathey. Extended depth of field through wave-front coding. *Appl. Opt.*, 1995. 1, 2, 3, 6, 7

[11] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019. 2

[12] Shay Elmalem, Raja Giryes, and Emanuel Marom. Learned phase coded aperture for the benefit of depth of field extension. *Opt. express*, 2018. 2, 3

[13] Yotam Gil, Shay Elmalem, Harel Haim, Emanuel Marom, and Raja Giryes. Monster: Awakening the mono in stereo. *arXiv preprint*, 2019. 2

[14] Joseph W Goodman. *Introduction to Fourier optics*. 2005. 4

[15] Harel Haim, Shay Elmalem, Raja Giryes, Alex M Bronstein, and Emanuel Marom. Depth estimation from a single image using deep learned phase coded mask. *TCI*, 2018. 2

[16] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 2007. 2

[17] Ralph Jacobson, Sidney Ray, Geoffrey G Attridge, and Norman Axford. *Manual of Photography*. 2000. 3

[18] Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *CVPR*, 2016. 2

[19] Lingbo Jin, Yubo Tang, Yicheng Wu, Jackson B Coole, Melody T Tan, Xuan Zhao, Hawraa Badaoui, Jacob T Robinson, Michelle D Williams, Ann M Gillenwater, et al. Deep learning extended depth-of-field microscope for fast and slide-free histology. *PNAS*, 2020. 2

[20] Sungkil Lee. Real-time depth-of-field rendering using point splatting on per-pixel layers. In *Comput Graph Forum*, 2008. 4

[21] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *TOG*, 2019. 2

[22] Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, and Mude Lin. Visually imbalanced stereo matching. In *CVPR*, 2020. 2

[23] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2, 4, 5

[24] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *CVPR*, 2020. 2

[25] Hajime Nagahara, Sujit Kuthirummal, Changyin Zhou, and Shree K Nayar. Flexible depth of field photography. In *ECCV*, 2008. 2, 3

[26] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nat. Methods*, 2020. 2

[27] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCVW*, 2017. 2, 5

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[29] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 2

[30] Yoav Shechtman, Steffen J Sahl, Adam S Backer, and WE Moerner. Optimal point spread function design for 3d imaging. *PRL*, 2014. 5, 6, 7

[31] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *TOG*, 2018. 2, 3, 7

[32] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *CVPR*, 2020. 2

[33] Yuichi Takeda, Shinsaku Hiura, and Kosuke Sato. Fusing depth from defocus and stereo with coded apertures. In *CVPR*, 2013. 2

[34] Ting-Chun Wang, Manohar Srikanth, and Ravi Ramamoorthi. Depth from semi-calibrated stereo and defocus. In *CVPR*, 2016. 2

[35] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *ICCP*, 2019. 2, 5, 7

[36] Yicheng Wu, Vivek Boominathan, Xuan Zhao, Jacob T Robinson, Hiroshi Kawasaki, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Freecam3d: Snapshot structured light 3d with freely-moving cameras. In *ECCV*, 2020. 2

[37] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. In *SIGGRAPH*. 2007. 7

[38] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, 2019. 2