Passive Snapshot Coded Aperture Dual-Pixel RGB-D Imaging

Bhargav Ghanekar
Rice University
Houston TX USA

Salman Siddique Khan
Rice University
Houston TX USA

Pranav Sharma
IIT Madras
Chennai India

Shreyas Singh
IIT Madras
Chennai India

Vivek Boominathan
Rice University
Houston TX USA

Kaushik Mitra
IIT Madras
Chennai India

Ashok Veeraraghavan
Rice University
Houston TX USA

Abstract

Passive, compact, single-shot 3D sensing is useful in many application areas such as microscopy, medical imaging, surgical navigation, and autonomous driving where form factor, time, and power constraints can exist. Obtaining RGB-D scene information over a short imaging distance, in an ultra-compact form factor, and in a passive, snapshot manner is challenging. Dual-pixel (DP) sensors are a potential solution to achieve the same. DP sensors collect light rays from two different halves of the lens in two interleaved pixel arrays, thus capturing two slightly different views of the scene, like a stereo camera system. However, imaging with a DP sensor implies that the defocus blur size is directly proportional to the disparity seen between the views. This creates a trade-off between disparity estimation vs. deblurring accuracy. To improve this trade-off effect, we propose CADS (Coded Aperture Dual-Pixel Sensing), in which we use a coded aperture in the imaging lens along with a DP sensor. In our approach, we jointly learn an optimal coded pattern and the reconstruction algorithm in an end-to-end optimization setting. Our resulting CADS imaging system demonstrates improvement of >1.5 dB PSNR in all-in-focus (AIF) estimates and 5-6% in depth estimation quality over naive DP sensing for a wide range of aperture settings. Furthermore, we build the proposed CADS prototypes for DSLR photography settings and in an endoscope and a dermoscope form factor. Our novel coded dual-pixel sensing approach demonstrates accurate RGB-D reconstruction results in simulations and real-world experiments in a passive, snapshot, and compact manner.

1. Introduction

Extracting 3D information from 2D images has widespread applications in several domains such as microscopy, medical imaging, and autonomous driving. Conventional methods like passive stereo, structured illumination [5, 24], photometric stereo [3], and time-of-flight sensing do an excellent job estimating depth maps from 2D images for several computer vision tasks. However, these 3D imaging methods have a bulkier form factor due to the use of multiple cameras/illumination sources, and their estimation accuracy suffers when they are used in systems with size restrictions.

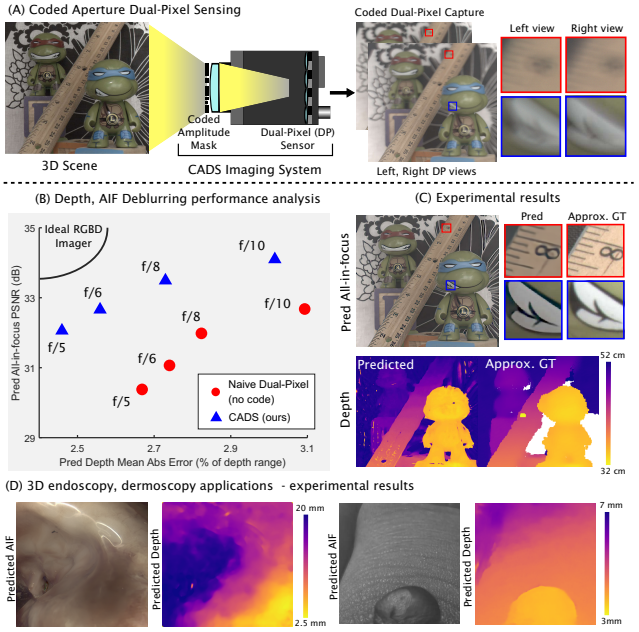

Figure 1. (A). We propose CADS - an imaging approach that leverages DP sensors and coded aperture masks for passive snapshot 3D imaging. (B) Coded DP sensing improves on naive dual-pixel sensing for simultaneous depth estimation and deblurring. Our CADS imaging prototype can recover accurate depth maps and all-in-focus images in (C) DSLR photography settings, as well as for (D) 3D endoscopy and dermoscopy. Read more on our website https://shadowfax11.github.io/cads/.

Monocular depth estimation has the potential to overcome these form factor restrictions. A popular monocular depth sensing strategy is coded aperture imaging, in which the aperture of the lens is coded either using an amplitude [16, 25] or a phase mask [31]. However, the quality of the depth maps and all-in-focus images obtained using these coded-aperture systems is still subpar compared to traditional methods like stereo or time-of-flight sensing. Recently, dual-pixel (DP) sensors have emerged as an accurate monocular depth sensing technology that overcomes these drawbacks due to their unique sensor design. These sensors are commonly found in smartphones and DSLR cameras.

Although DP sensors were primarily developed for fast and accurate auto-focusing in consumer-grade cameras, there have been numerous works [15, 20, 22, 32] that have tried to use these sensors as single-shot depth and intensity imaging sensors. Specifically, because of the unique microlens array arrangement over the photodiodes, these sensors can capture two blurry views of the same scene in a single capture. Moreover, the disparity between these two views is directly related to the defocus blur size. The small disparity in the views can be used for depth estimation while deblurring the defocus blur can provide us with the all-in-focus (AIF) image estimate. However, existing DP AIF estimation works are inherently limited by poor conditioning of the DP defocus blur. The inability to deblur with high fidelity then severely restricts their depth-of-field.

To improve the depth-of-field of dual-pixel cameras while still maintaining their ability to capture depth information, we introduce a novel sensing strategy called CADS - **C**oded **A**perture **D**ual-pixel **S**ensing. CADS uses a learned amplitude mask (code) in the aperture plane of a camera equipped with a DP sensor to improve the conditioning of the DP defocus blur while maintaining the disparity estimation ability from the two views. We use a two-stage, end-to-end learning framework to simultaneously learn the optimal coded mask pattern and the optimal neural network weights that can produce an accurate depth map and deblurred all-in-focus image estimates from dual-pixel measurements for a wide range of aperture settings (see Fig. 2).

Compared to conventional DP sensing (with no coded aperture), the learned coded dual-pixel sensing system provides a better AIF performance, indicated through a consistent $\geq$ 1.5dB PSNR improvement in the quality of AIF images and a 5-6% improvement in the quality of depth maps for a wide range of aperture settings, allowing a better disparity and AIF quality trade-off. We verify the efficacy of CADS through various simulations and real-world experiments. CADS provides a mechanism for high-fidelity depth and intensity imaging in a small form factor. We demonstrate its benefits in two relevant real-world applications - 3D endoscopy and extended depth-of-field (EDOF) dermoscopy. For both, we demonstrate promising real-world experiments using proof-of-concept prototypes built in the lab, see Fig. 1. In summary, our contributions are:

- We propose coded aperture dual-pixel sensing, called CADS, that significantly improves the trade-off between quality of disparity and AIF estimates from dual-pixel sensors.
- At the core of our method is the proposed differentiable coded dual-pixel sensing model that allows end-to-end learning of the coded aperture mask specific to the dual-pixel sensor geometry.
- We highlight the efficacy of our learned code design through various simulations and real-world experiments.
- We show the effectiveness of coded aperture dual-pixel systems for close-range, compact RGB-D imaging by building a coded dual-pixel 3D endoscope and a coded dual-pixel dermoscope.

2. Related works

2.1. Dual-Pixel sensing

Dual-pixel (DP) sensors were initially used for fast autofocus [12, 28], primarily using the fact that in-focus (out-of-focus) scene points show zero (non-zero) disparity. Since then, they have been used for several tasks. Exploiting the disparity cue, [21] demonstrated the use of DP sensors for reflection removal. [30] used classical stereo algorithms for disparity estimation, to simulate shallow depth-of-field. Since then, there have been several works on disparity or depth estimation using dual-pixel sensors [8, 15, 20, 22, 32]. [35] combines dual-pixel data with binocular stereo data to predict accurate disparity maps. Most of the methods use optimization techniques [22, 32] or leverage deep networks in supervised [8, 15, 20] or self-supervised [15] settings. Apart from disparity estimation, the authors in [14] show normal estimation of human faces as well. [20, 32] have been shown on unidirectional disparity only. [15] addressed this and came up with a self-supervised method for DP images showing bidirectional disparity. There also have been numerous works on deblurring DP captures [1, 2, 20, 32, 33], which too are based on optimization methods or deep neural networks. In contrast to the above works, which mainly focus on post-processing conventional dual-pixel captures, we propose a novel modification to DP sensing strategy by adding a coded mask in the aperture plane that allows us to achieve better disparity and AIF quality trade-off for dual-pixel sensing.

2.2. Coded aperture imaging

Multiplexing of incoming light using a coded mask in the aperture plane has been used for numerous imaging applications over the years. Levin *et al*. [16] used a designed amplitude mask in the aperture plane for estimating a depth map and an all-in-focus image from a single defocused im-

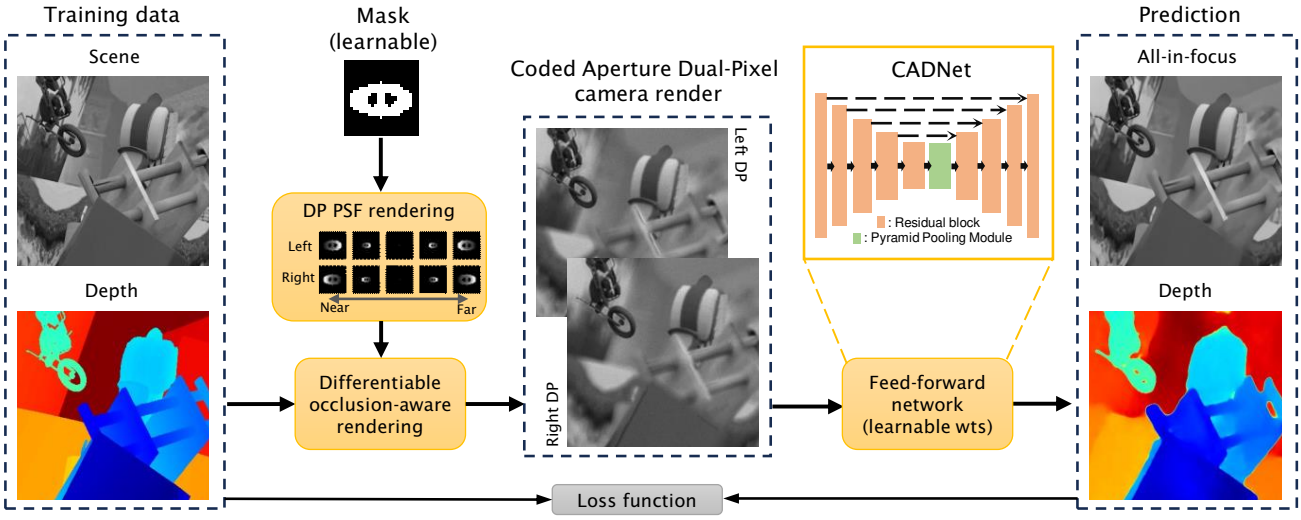

Figure 2. **Pipeline for Coded Aperture Dual-Pixel Sensing (CADS) and corresponding 3D scene estimation.** We perform end-to-end (E2E) optimization on simulated data, to learn an optimal amplitude mask and neural network weights for predicting deblurred all-in-focus (AIF) images and depth maps from coded dual-pixel captures. Our end-to-end learned system provides the best trade-off between depth estimation and AIF quality.

age. Veeraraghavan *et al*. [29] used amplitude-coded aperture patterns for improved light-field imaging. Zhou *et al*. [38] use a classical genetic optimization scheme to obtain coded aperture designs for estimating depth from coded defocus pairs. Since the design of the code or the mask pattern plays an important role in the downstream task, recently, there have been numerous works on using end-to-end learning schemes for designing the optimal coded mask patterns. Shedligeri *et al*. [25] learn the optimal amplitude mask pattern for estimating depth and all-in-focus image from a single defocused capture and report an improvement over the heuristically designed mask pattern proposed in [16]. A phase mask can also be used to introduce coding in the aperture plane. Recently, works like [9, 31] have used end-to-end learning to design the optimal phase mask profiles for improved 3D imaging. The proposed CADS system also uses an end-to-end learned amplitude mask in the aperture plane. Compared to existing coded aperture works, we use this amplitude mask for a DP sensor and are the first to do so. The use of a learned code in a DP setting allows the recovery of high-fidelity depth and AIF for a wide range of depth compared to standard coded-aperture imaging.

3. Coded Aperture Dual-pixel Sensing (CADS)

3.1. Coded dual-pixel image formation model

In a conventional image sensor, each pixel consists of one photodiode and a microlens to gather incoming light. In a dual-pixel (DP) sensor, each pixel consists of one microlens that covers two photodiodes, say left and right. Such a configuration enables all the left(right) photodiodes to receive light specifically from the right(left) half of the lens (see Fig. 3A). Thus with a DP sensor, we obtain two images of the same scene from slightly different viewpoints, which are commonly referred to as the left and right DP images. Scene points that are in-focus are exactly aligned in both views, while scene points that are out-of-focus show positive or negative disparity when they are farther or nearer to the lens respectively.

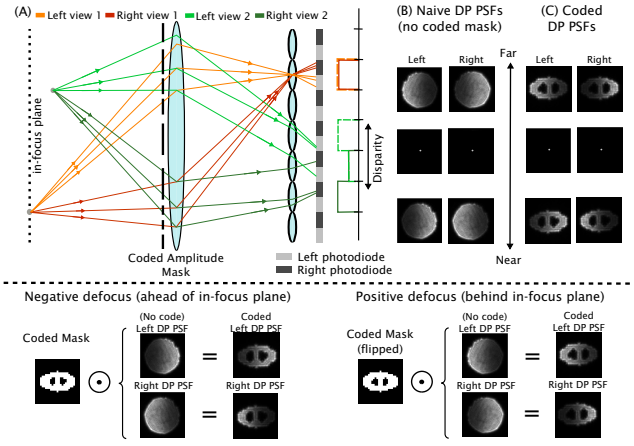

Figure 3. **CADS PSF formation model.** (A) With a DP sensor, scene points that are out-of-focus show disparity when comparing between left and right views. (B) This disparity is bi-directional, depending on scene depth relative to the in-focus plane. (C) When a coded amplitude mask is added to the imaging lens, the DP PSFs take the shape of the mask pattern. (D) Coded DP PSFs are a Hadamard product of the naive (no code) DP PSFs and the mask pattern.

However, this disparity comes with an undesirable defocus blur effect of defocus blur. In a DP imaging system, the

defocus blur size is directly proportional to the disparity. To have good, accurate depth measurements, one needs to have a large disparity range between the closest and the farthest scene depths, but this comes at the cost of highly defocused, blurry images, leading to poor AIF image estimation.

To improve upon all-in-focus recovery while preserving the depth sensing capability of DP sensors, we propose the use of coded amplitude masks alongside DP sensors.

When a coded (amplitude) mask M is added to an imaging system/camera with DP sensor, the *coded* DP PSFs can be expressed as a Hadamard (element-wise) multiplication of the naive DP PSFs with the coded mask. This is illustrated in Fig. 3. More formally, the left and right DP PSFs ($h_z^{L,C}$ and $h_z^{R,C}$ respectively) can be expressed as

$$h_z^{L,C} = \underset{z>z_f}{\text{flip}}\,(M) \odot h_z^L, \quad h_z^{R,C} = \underset{z>z_f}{\text{flip}}\,(M) \odot h_z^R \tag{1}$$

where h_z^L, h_z^R are the left and right view PSFs in the naive DP case (with no mask), M is simply the mask pattern (resized to the blur kernel size), $\text{flip}_{z>z_f}(.)$ performs horizontal and vertical flipping when the point source is in positive defocus and is an identity function otherwise.

3.2. CADS framework

Our framework for Coded Aperture DP Sensing (or CADS) consists of two components -

- **Coded dual-pixel image simulator**. Given scene intensity map (RGB or grayscale), scene depth map, coded mask pattern, and the naive (no code) DP PSFs, we use a differentiable rendering pipeline to generate the left, right CADS images of a scene.
- **CADNet**. Given the left, right CADS images of a scene, we propose to use a neural network (which we call CADNet) that will take these images as inputs and produce a normalized defocus map and a deblurred all-in-focus (AIF) image as outputs.

We learn the two components in an end-to-end manner.

3.2.1 Coded dual-pixel image simulator

Mask-to-PSF generation. In our CADS framework, we parameterize our mask pattern M with a continuous-valued map $\theta_C(x, y)$ as done in [25], where $M_{\theta_C}(x, y) = \sigma(\alpha * \theta_C(x, y))$, where $\sigma(.)$ is the sigmoid function. α is a scalar temperature factor that controls mask transparency for smooth learning[1]. For realistic modeling of the naive (no-code) DP PSFs h_z^L, h_z^R, we use the formulation provided in Abuolaim *et al.* [2]. We generate left and right naive DP PSFs for 21 depth planes, with signed blur sizes ranging from -40 pixels to $+40$ pixels. More details regarding this can be found in the supplementary. Given M_{θ_C}, h_z^L, h_z^R, we use Eq. 1 to generate the coded left, right DP PSF z-stack i.e. $h_z^{L,C}$ and $h_z^{R,C}$ for all the 21 depth planes.

Coded DP image rendering. We adopt a multi-plane image (MPI) representation of the scene, where the 3D scene is modelled as several discrete depth planes. Given $h_z^{L,C}$ and $h_z^{R,C}$, the coded DP left, right images I_L, I_R of a 3D scene can be expressed as follows -

$$I_L = \sum_z h_z^{L,C} * s_z, \quad I_R = \sum_z h_z^{R,C} * s_z, \tag{2}$$

where s_z is the scene intensity at depth z, and $*$ is the 2D convolution operator. To remove artifacts at the edges of the MPI depth layers, we adopt the differentiable occlusion-aware modifications to Eqn. 2 from Ikoma *et al.* [11], along with a minor modification. See supplementary for details.

3.2.2 CADNet for defocus and AIF estimation

Recovering the depth map and the deblurred all-in-focus (AIF) image from blurry measurements can be solved by formulating a convex optimization problem [32]. However, such optimization methods are usually slow. Thus, we employ a neural network (referred to as CADNet), to recover depth and AIF images of the scene. CADNet architecture is based on the U-Net architecture [23] that applies multi-resolution feature extraction with skip connections between encoder and decoder blocks of the same scale. We modify the basic U-Net architecture (illustrated in Fig. 2) to have a residual block [10, 36] with two convolution layers of kernel size 3×3 at each scale. In addition, we append a pyramid pooling module (PPM) [37] at the lowest scale. A pixel-shuffle convolution layer [26] is used in the encoder for downsampling, while bilinear upsampling is used in the decoder. Our neural network is trained entirely on a synthetic dataset of coded DP images rendered from the FlyingThings3D dataset.

CADNet takes as input the blurry measurements and outputs - (1) a normalized defocus map $\hat{D}_N$, ranging from -1 to +1, and (2) the deblurred all-in-focus (AIF) image $\hat{Y}$ of the scene. The normalized defocus map $\hat{D}_N$ is converted to actual physical defocus blur size by scaling it as $\hat{D} = \hat{D}_N D_{max}/p$, where p is the pixel pitch of the DP sensor, and D_{max} being the maximum blur size (in px) that CADNet is trained to handle. This defocus value is related to the imaging depth as follows [8]

$$D(z) = \frac{Lf}{1 - f/g}\left(\frac{1}{g} - \frac{1}{z}\right), \tag{3}$$

where L is the diameter of the imaging lens, f is the focal length, g is the in-focus distance, and z is the depth of the object being imaged. Thus, given information about the imaging system we can readily calculate the depth from the CADNet defocus map output using Eq. 3. For the AIF

[1] We gradually increase α over iterations to ensure the mask is learnt smoothly to be binary (0 or 1). See Supplementary for more details.

output channel from CADNet, we output normalized AIF images ranging from 0 to 1. We train two variants of CADNet, namely CADNet-Mono and CADNet-RGB. CADNet-Mono takes 2 input channels - the coded left, right DP images from a single, monochrome (grayscale) color, and the AIF outputs are single channel only. CADNet-RGB takes 6 input channels - the coded left, right DP images for each of the red, green, and blue channels, and the AIF output is a 3-channel RGB image. We train these two variants because real-world DP sensors present in the Pixel 4 and Canon EOS Mark IV DSLR are single-channel (green only) and full-frame (RGB) respectively.

Loss functions. We jointly optimize for the mask pattern parameters (θ_C) and CADNet weights (θ_D) by minimizing the following loss function

$$\mathcal{L} = \mathcal{L}_{AIF} + \mathcal{L}_{defocus} + \mathcal{L}_{mask}, \tag{4}$$

$$\mathcal{L}_{AIF} = \beta_1 \|\hat{Y} - Y_{gt}\|_1 + \beta_2 \|\nabla\hat{Y} - \nabla Y_{gt}\|_1, \tag{5}$$

$$\mathcal{L}_{defocus} = \beta_3 \|\hat{D} - D_{gt}\|_1 + \beta_4 \|\nabla\hat{D} - \nabla D_{gt}\|_1, \tag{6}$$

$$\mathcal{L}_{mask} = \beta_5 \text{Relu}\left(0.5 - \frac{\sum_{x,y} M_{\theta_C}}{\sum_{x,y} M_{open}}\right) \tag{7}$$

where $\|.\|_1$ is the L_1 loss, $\nabla(.)$ is 2D gradient operator. $\hat{Y}$ and $\hat{D}$ are predicted AIF and defocus maps, while Y_{gt} and D_{gt} are groundtruth AIF and defocus maps. The last term is a mask pattern regularizer, where M_{open} is the open aperture mask pattern. We set $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (1, 0.5, 1, 0.5, 1, 10^3)$, with $\beta_5 >> 1$ to enforce a strict transmission constraint ($\geq 50\%$ light efficiency).

3.2.3 Implementation details

For training the CADS framework, we used simulated data generated using the forward model (discussed in Section 3.2) on FlyingThings3D dataset scenes[17]. 20k scenes are split $80\%-20\%$ into training and validation sets. We rescale the disparity maps for each scene to convert them to depth maps having range $\mathcal{D}_T$ = [32 mm, 76 mm]. During training of our coded DP simulator, camera parameters were set at $f = 4.38$ mm, $L = f/1.73$, $g = 45$ mm, ensuring that depth range $\mathcal{D}_T$ corresponded to blur sizes of ≤ 40 pixels (based on Eq. 3). During testing on simulated data, we set camera parameters to match our real-world setting with focal length $f = 50$ mm, aperture $L = f/4$, in-focus distance $g = 40$ cm, with a depth range $\mathcal{D}_V$ = [32 cm, 52 cm]. The signed blur sizes for $\mathcal{D}_V$ again corresponded to ≤ 40 pixels (pixel pitch, $p = 10.72\mu m$). Using the loss function described in Eqn. 4, we perform end-to-end training, learning weights for CADNet and learning a coded mask pattern with light efficiency $>= 50\%$. To account for light loss due to coded mask, we add appropriate heteroscedastic noise [6] to simulated captures. We train the end-to-end framework using the Adam optimizer. We initialize the coded aperture mask with an open aperture pattern of size 21x21 pixels. More details are in the supplementary.

4. Experiments and results

4.1. Comparison with naive standard and dual pixel sensing

As a result of end-to-end training, we learn an optimal coded mask pattern that better conditions the coded DP blurry measurements, thus helping in good depth and AIF recovery. The learned mask and the corresponding left, right coded DP PSFs as a result of our end-to-end training have been depicted in Figs. 2, 3. We refer the readers to the Supplementary for further illustrations and explanations. In this section, we show that our learned coded-aperture DP sensing is in fact an improvement over naive monocular standard and dual-pixel sensing in terms of both AIF and depth estimation quality. We use PSNR, SSIM, and LPIPS measures[34] for the evaluation of AIF quality. We use Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and depth estimation accuracy within a threshold margin ($\delta^1 < 1.05^1$) for the evaluation of depth prediction quality (see Supplementary for definitions).

Depth Prediction			
Imager	RMSE(mm)↓	MAE(mm)↓	δ^1↑
Naive SP	48.51	29.37	0.688
Naive DP	13.66	5.51	0.967
CADS	12.57	5.15	0.972
AIF Prediction			
Imager	PSNR(dB)↑	SSIM↑	LPIPS ↓
Naive SP	27.69	0.790	0.427
Naive DP	29.72	0.832	0.381
CADS	31.20	0.865	0.337

Table 1. **Comparison with naive standard and DP - simulation.** Scenes with 20 cm depth range, rendered with f/4 aperture. Our proposed CADS shows the best performance for joint depth and AIF recovery. Red highlights best, orange highlights second best. (SP: Standard-pixel, DP: Dual-pixel)

Simulation results. We simulate measurements using the FlyingThings3D dataset and report the performance of CADS on it. We also compare CADS with two other imaging systems, namely, (a) Naive Std. Pixel - imaging system with a conventional standard pixel sensor and no aperture coding, (b) Naive Dual-pixel - imaging system with a dual-pixel sensor and no aperture coding. We trained a separate CADNet for each imaging system. Our simulation results are shown in Table 1 and Fig. 4. When compared to the naive standard pixel scenario, both naive dual-pixel and CADS improve the deblurring and depth estimation accuracy significantly. Furthermore, CADS shows a 1.5dB gain

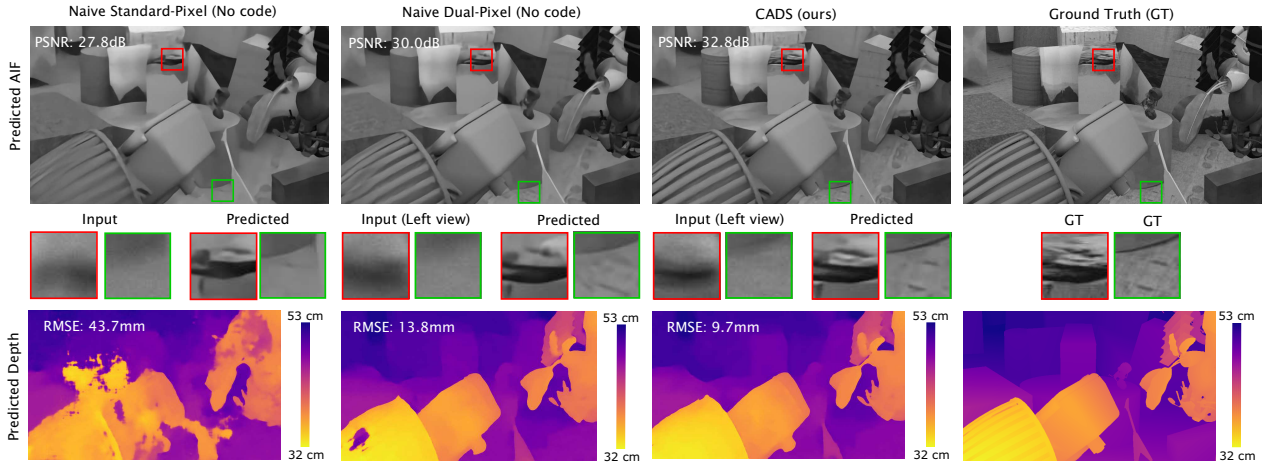

Figure 4. **Comparison with naive standard and DP - simulation.** DP sensing methods provide high-fidelity depth and all-in-focus (AIF) prediction compared to standard pixel sensing. Among DP sensing methods, CADS offers the best AIF and depth estimation quality observed through sharper AIF and cleaner depth.

in PSNR and a 5-8% gain in depth prediction accuracy over naive dual-pixel.

Real results. We perform real-world experiments by building a prototype imaging system capable of capturing coded dual-pixel images. We use the Canon EOS 5D Mark IV DSLR with a Yongnuo 50 mm lens. The Canon EOS Mark IV sensor is an RGB, full-frame dual-pixel sensor, with every pixel in the R-G-G-B Bayer pattern being dual-pixels. Our coded mask pattern of diameter $L = f/4 = 12.5$ mm was printed on a transparency sheet and placed in the aperture plane of the Yongnuo lens. More details about the setup can be found in the supplementary. For accurate reconstruction, we render FlyingThings3D scenes with the real, captured PSFs, and fine-tune CADNet weights for 30 epochs with real-world captured PSFs. More details regarding fine-tuning procedure can be found in the supplementary. Fig. 5 shows real-world results with our imaging prototype. Our proposed CADS system provides significantly better AIF estimates while preserving the depth-sensing ability of dual-pixel sensors.

4.2. Performance of CADS for various aperture sizes

CADS offers a better trade-off between depth and AIF quality compared to a naive dual-pixel sensor for various aperture sizes as shown in Fig. 1. To establish this, we evaluate the performance of CADS for various aperture sizes ranging from $f/4$ to $f/10$. We test this out on the same FlyingThings3D scenes. Apart from comparing CADS against naive dual-pixel sensing (Naive DP), we also compare it against coded and no-code variants of standard pixel sensor, referred to as Coded and Naive SP respectively. Table 2 shows depth estimation and AIF deblurring results on our simulated FlyingThings3D data, where the imaging system

Depth Prediction MAE (mm) ↓				
Imaging System	Aperture size (f/#)			
	f/4	f/6	f/8	f/10
Naive SP	26.64	26.45	26.26	24.39
Coded SP	21.37	22.18	21.83	22.68
Naive DP	5.44	5.48	5.65	6.19
CADS	4.99	5.12	5.46	6.03
AIF Prediction PSNR (dB) ↑				
Imaging System	Aperture size (f/#)			
	f/4	f/6	f/8	f/10
Naive SP	27.66	29.16	30.08	30.87
Coded SP	28.97	30.28	31.16	31.90
Naive DP	29.54	31.07	31.98	32.68
CADS	31.21	32.66	33.49	34.10

Table 2. **Performance under varying f-number**. The proposed CADS performs the best passive snapshot depth and all-in-focus estimation across different aperture sizes. Red highlights the best method, and orange highlights the second best.

parameters are the same as mentioned before, with objects being placed in the [32 cm, 53 cm] range. With decreasing aperture size (increasing f-number), AIF prediction accuracy increases while depth prediction gets worse. Our proposed E2E CADS system consistently shows the best performance across all apertures and improves the defocus-disparity trade-off.

4.3. Comparison with existing dual-pixel works

We compare the performance of CADS with existing learning-based dual-pixel sensing works, namely DPDNet[2], DDDNet[20], Xin *et al*. [32], Punnapurath *et al*. [22] and Kim *et al*. [15]. Since existing works were developed for naive (no-code) dual-pixel sensors, we use the naive DP PSF blur to simulate captures for evaluation. For evaluating CADS, we render using the coded-aperture DP

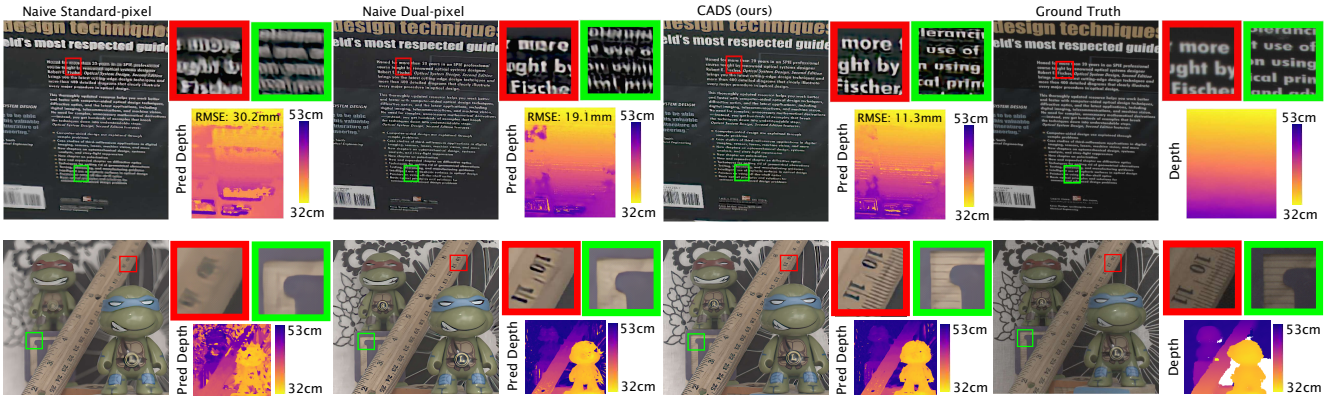

Figure 5. **Comparison with naive standard and DP - real data.** From left to right we show AIF and depth predictions by naive standard-pixel, naive dual-pixel and CADS, respectively, along with ground truth. CADS gives the best depth and AIF predictions. GT for AIF is obtained by capturing the scene with a $f/22$ aperture. A coarse GT depth map is obtained using the Intel RealSense sensor. Zoom in for best viewing.

Method	AIF PSNR	Disp. AI(1)
DPDNet [2]	24.34	N.A.
DDDNet [20]	21.35	0.2769
Xin *et al*. [32]	18.13	0.3018
Punnappurath *et al*. [22]	N.A.	0.1940
Kim *et al*. [15]	N.A.	0.2866
Naive DP (ours)	29.72	0.0190
CADS (ours)	31.20	0.0177

Table 3. **Comparison with existing DP methods.** CADS offers the best AIF and disparity estimation quality on our simulated dataset. Red highlights best, orange highlights second best.

PSF. We use a validation subset of 2k images from the FlyingThings3D dataset to evaluate. The results are shown in Table 3. Most existing works estimate disparity which is related to defocus map by an unknown scale. For evaluation, we use the affine invariant version of MAE (AI(1)) for disparity estimation quality [8] and PSNR for AIF quality[2]. Note that we use the normalized defocus map output by CADNet for this comparison instead of converting it to a depth map. Works that were designed for unidirectional disparity [20, 32] show poor results when tested on simulated scenes with bidirectional disparity between the DP images. Our proposed CADS outperforms existing methods on the simulated FlyingThings3D dataset. Moreover, CADNet trained on naive DP blurs, referred to as Naive DP in Table 3, also outperforms existing methods. We also perform a comparison on rendered images from NYUv2 dataset[27] and report the performance in the supplementary.

5. Applications

We build two prototype CADS systems for 3D endoscopy and 3D dermoscopy and show real-world endoscopy and dermoscopy results using these prototypes.

[2]See Supplementary for metric definitions

5.1. CADS Endoscopy

3D endoscopy is potentially useful for surgical navigation, tumor/polyp detection, etc [4, 18]. Stereo-endoscopes have been used for obtaining 3D information of the scene, but they are not compact. 3D endoscopy solutions have been attempted by using SfM methods [19], but they lack accuracy due to non-rigid nature of human tissue and organs. Structured illumination methods have also been used for 3D endoscopy [7]; however, this requires close control over illumination, which is difficult. Thus, there is a need for developing passive snap-shot depth sensing in endoscopy.

CADS can potentially be a mechanism to achieve passive, snapshot, per-frame 3D endoscopy where depth maps as well as AIF images are simultaneously predicted. We built a coded aperture dual-pixel endoscopic sensing prototype as shown in Fig. 6. We use Canon EOS Mark IV DSLR with a Yongnuo 50 mm lens as our dual-pixel camera and a Karl Storz Rubina 10mm Endoscope. The camera and the scope are coupled by aligning both their optical axis and placing the camera lens close to the endoscope eyepiece. We put our coded aperture pattern on a 3D-printed circular aperture and placed it in the middle of the scope and the DSLR lens. We focus our lens such that the in-focus plane is 2 cm away from the other end of the endoscope. We capture PSFs for negative defocus and fine-tune only for the negative defocus region. To account for the spatial variance in the PSF, we randomly sample PSFs across the field of view for training. Details regarding the setup and PSF calibration are in the supplementary.

To emulate realistic endoscopy scenes, we image lamb chops tissue samples. Depth and AIF reconstruction results (image size ∼800x800) for such scenes are shown in Fig. 6. Depth maps are shown after 21x21 median filtering. Despite being finetuned on FlyingThings3D dataset, CADNet is still able to recover sharp AIF and clean depth maps.

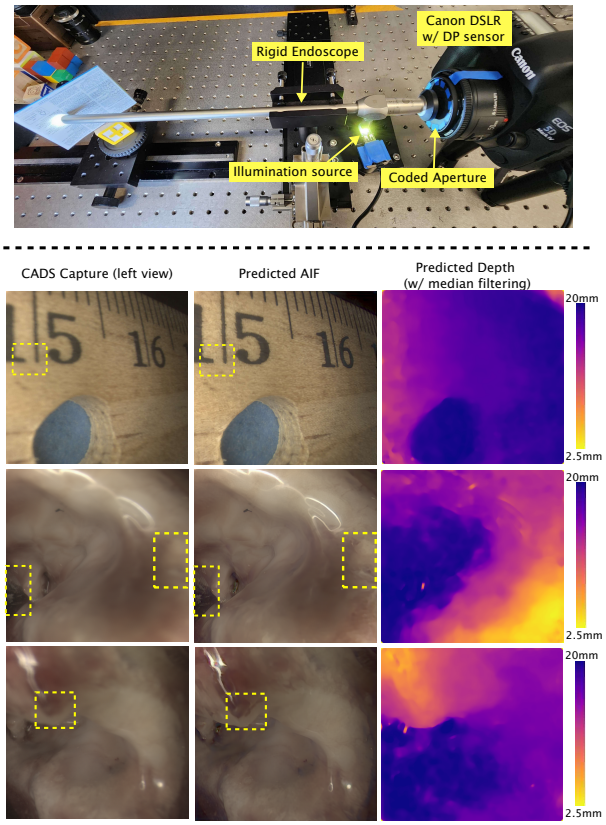

Figure 6. **CADS endoscopy results.** (Top) Proof-of-concept 3D endoscopy setup built using CADS. (Bottom) CADS allows us to recover sharp AIF and accurate depth for various scenes (∼800x800 px recon.). Zoom-in on yellow boxes for best results.

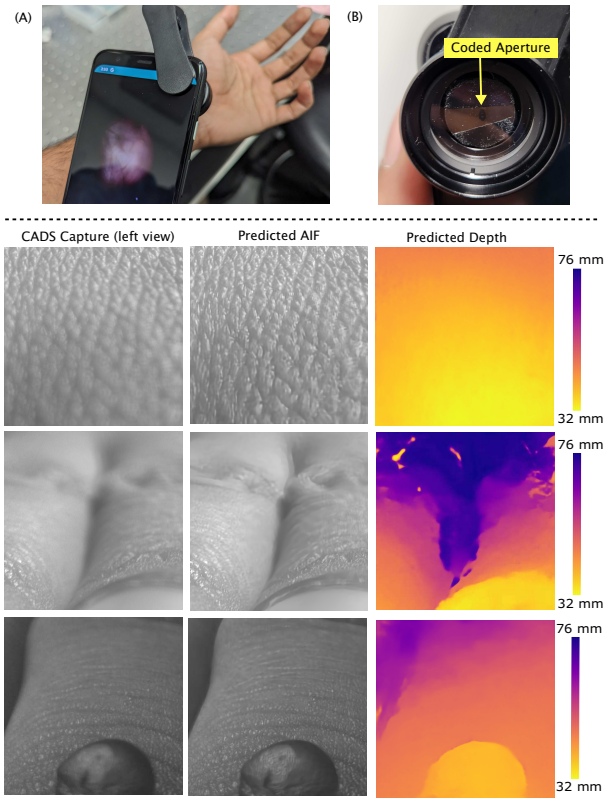

Figure 7. **CADS dermoscopy results.** (Top)(A)Prototype CADS dermoscope using Pixel 4 and 12× macro lens.(B)The coded aperture is placed in between the smartphone lens and camera lens.(Bottom)Example AIF, depth reconstructions (∼450x450 px)

5.2. CADS Dermoscopy

Dermoscopes are imaging systems used in clinical settings to view skin lesions. Existing non-contact dermoscopes suffer from small depth-of-field and can only provide 2D images. [13] collect a focal stack to overcome this drawback. However, this is time-consuming, prone to motion blur, and doesn't provide depth information. We build an in-house smartphone-based CADS dermoscope, using a Pixel 4 smartphone that is equipped with a DP sensor and 12x macro lens. We finetuned our CADNet model on measurements simulated from FlyingThings3D images using the captured spatially-varying PSFs. Fig. 7 shows the visual results for depth and AIF predictions from our in-house dermoscope estimated using the fine-tuned CADNet on captures of a human subject's hand. The use of CADS in a dermoscope offers high-fidelity AIF and depth maps for a wide range of depth despite being finetuned on FlyingThings3D dataset.

6. Conclusions

Dual-pixel sensors have emerged as a novel compact depth-sensing modality with widespread use in consumer-grade cameras. However, existing dual-pixel depth sensing methods suffer from limited depth of field (DOF) due to an inherent trade-off between the quality of depth maps and the ease of deblurring. Moreover, the majority of existing works fail for bidirectional defocus/disparity. To overcome these drawbacks, we use ideas from coded aperture imaging to develop a novel, single-shot 3D sensing strategy called CADS - Coded Aperture Dual-pixel Sensing. More specifically, we improve the conditioning of the dual-pixel defocus blur by introducing an end-to-end learned amplitude mask in the aperture plane. Introducing the learned mask improves the quality of deblurred all-in-focus images and depth maps for a wide range of depth. We demonstrate the efficacy of our proposed systems for three different applications: DSLR photography, 3D endoscopy, and extended DOF dermoscopy. Promising results on real data collected using proof-of-concept hardware indicate CADS can be used for applications beyond photography like medical imaging where form-factor plays a pivotal role. In the future, it would be interesting to replace amplitude masks with phase masks [31] that can improve the light efficiency and as a result, the SNR of the system.

Acknowledgements. This work was supported by NSF: IIS-1730574, NIH: R01DE032051-01. K.M. acknowledges NSF/IITM Pravartak Technologies Foundation and funding from the Department of Science and Technology, India.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 111–126. Springer, 2020. 2

[2] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2289–2298, 2021. 2, 4, 6, 7

[3] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of computer vision*, 72:239–257, 2007. 1

[4] Robert Bickerton, Abdul-Karim Nassimizadeh, and Shahzada Ahmed. Three-dimensional endoscopy: The future of nasoendoscopic training. *The Laryngoscope*, 129 (6):1280–1285, 2019. 7

[5] Kim L Boyer and Avinash C Kak. Color-encoded structured light for rapid active ranging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):14–28, 1987. 1

[6] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008. 5

[7] Ryo Furukawa, Hiroki Morinaga, Yoji Sanomura, Shinji Tanaka, Shigeto Yoshida, and Hiroshi Kawasaki. Shape acquisition and registration for 3d endoscope based on grid pattern projection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 399–415. Springer, 2016. 7

[8] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T. Barron. Learning single camera depth estimation using dual-pixels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 4, 7

[9] Harel Haim, Shay Elmalem, Raja Giryes, Alex M Bronstein, and Emanuel Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4(3):298–310, 2018. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[11] Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2021. 4

[12] Jinbeum Jang, Yoonjong Yoo, Jongheon Kim, and Joonki Paik. Sensor-based auto-focusing system using multi-scale feature extraction and phase correlation matching. *Sensors*, 15(3):5747–5762, 2015. 2

[13] Lennart Jütte, Zhiyao Yang, Gaurav Sharma, and Bernhard Roth. Focus stacking in non-contact dermoscopy. *Biomedical physics & engineering express*, 8(6):065022, 2022. 8

[14] Minjun Kang, Jaesung Choe, Hyowon Ha, Hae-Gon Jeon, Sunghoon Im, In So Kweon, and Kuk-Jin Yoon. Facial depth and normal estimation using single dual-pixel camera. In *European Conference on Computer Vision*, pages 181–200. Springer, 2022. 2

[15] Donggun Kim, Hyeonjoong Jang, Inchul Kim, and Min H. Kim. Spatio-focal bidirectional disparity estimation from a dual-pixel image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5023–5032, 2023. 2, 6, 7

[16] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26 (3):70–es, 2007. 2, 3

[17] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 5

[18] Kosuke Nomura, Daisuke Kikuchi, Mitsuru Kaise, Toshiro Iizuka, Yorinari Ochiai, Yugo Suzuki, Yumiko Fukuma, Masami Tanaka, Yosuke Okamoto, Satoshi Yamashita, et al. Comparison of 3d endoscopy and conventional 2d endoscopy in gastric endoscopic submucosal dissection: an ex vivo animal study. *Surgical endoscopy*, 33:4164–4170, 2019. 7

[19] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical image analysis*, 71:102058, 2021. 7

[20] Liyuan Pan, Shah Chowdhury, Richard Hartley, Miaomiao Liu, Hongguang Zhang, and Hongdong Li. Dual pixel exploration: Simultaneous depth estimation and image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2021. 2, 6, 7

[21] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1556–1565, 2019. 2

[22] Abhijith Punnappurath, Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Modeling defocus-disparity in dual-pixel sensors. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020. 2, 6, 7

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[24] Manish Saxena, Gangadhar Eluru, and Sai Siva Gorthi. Structured illumination microscopy. *Advances in Optics and Photonics*, 7(2):241–275, 2015. 1

[25] Prasan A Shedligeri, Sreyas Mohan, and Kaushik Mitra. Data driven coded aperture design for depth recovery. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 56–60. IEEE, 2017. 2, 3, 4

[26] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[27] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 7

[28] Przemysław Śliwiński and Paweł Wachel. A simple model for on-sensor phase-detection autofocusing algorithm. *Journal of Computer and Communications*, 1(06):11, 2013. 2

[29] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69, 2007. 3

[30] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2

[31] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019. 2, 3, 8

[32] Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T Barron, Pratul P Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. Defocus map estimation and deblurring from a single dual-pixel image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2021. 2, 4, 6, 7

[33] Yan Yang, Liyuan Pan, Liu Liu, and Miaomiao Liu. K3dn: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13272, 2023. 2

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[35] Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne, Sean Fanello, and Rahul Garg. Du 2 net: Learning depth estimation from dual-cameras and dual-pixels. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 582–598. Springer, 2020. 2

[36] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 4

[37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4

[38] Changyin Zhou, Stephen Lin, and Shree Nayar. Coded aperture pairs for depth from defocus. In *2009 IEEE 12th international conference on computer vision*, pages 325–332. IEEE, 2009. 3