Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

FlatNet3D: intensity and absolute depth from single-shot lensless capture

Dhruvjyoti Bagadthey,[1,†] Sanjana Prabhu,[1,†] Salman S. Khan,[1,*] D Tony Fredrick,[1] Vivek Boominathan,[2] Ashok Veeraraghavan,[2] and Kaushik Mitra[1]

[1] Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India
[2] Rice University, Houston, Texas 77005, USA
*Corresponding author: sk39@smail.iitm.ac.in

Received 8 June 2022; revised 8 September 2022; accepted 14 September 2022; posted 14 September 2022; published 28 September 2022

Lensless cameras are ultra-thin imaging systems that replace the lens with a thin passive optical mask and computation. Passive mask-based lensless cameras encode depth information in their measurements for a certain depth range. Early works have shown that this encoded depth can be used to perform 3D reconstruction of close-range scenes. However, these approaches for 3D reconstructions are typically optimization based and require strong hand-crafted priors and hundreds of iterations to reconstruct. Moreover, the reconstructions suffer from low resolution, noise, and artifacts. In this work, we propose *FlatNet3D*—a feed-forward deep network that can estimate both depth and intensity from a single lensless capture. FlatNet3D is an end-to-end trainable deep network that directly reconstructs depth and intensity from a lensless measurement using an efficient physics-based 3D mapping stage and a fully convolutional network. Our algorithm is fast and produces high-quality results, which we validate using both simulated and real scenes captured using PhlatCam.

https://doi.org/10.1364/JOSAA.466286

1. INTRODUCTION

Depth estimation from images is an important vision problem that has its applications in robotics, computer aided diagnosis, and autonomous systems with many of these systems having strict form-factor and weight constraints. This makes the use of existing depth estimation techniques like time-of-flight (ToF) sensors, structured light, or stereo camera systems difficult since these systems have a larger form-factor and are heavier by design. To overcome these challenges, the development of miniature light-weight 3D sensing cameras is imperative.

Mask-based lensless cameras have emerged as an ideal candidate for light-weight, small form-factor imaging systems. These cameras replace the lens of a traditional imaging system with an amplitude mask [1] or a phase mask [2,3] and computation. This results in systems in which the need for the lens, which is a major contributor toward size and weight, is eliminated. Moreover, given that a lensless imaging system is no longer restricted due to the rigidity imposed by the focal-length constraint of lenses, it can have a flexible form-factor [4]. Thus, being able to do 3D imaging using lensless cameras will allow us to have ultra-thin, light-weight 3D sensors.

In this work, we focus on mask-based lensless cameras. Passive mask-based lensless cameras can have an ultra-thin form-factor ($\sim 1-2$ mm), large rectilinear field of view, and depth discriminating characteristics, and they do not require coherence [1,2,5–8]. These properties make mask-based lensless cameras more attractive than existing coherent diffractive lensless imaging and incoherent light field imaging especially for imaging applications where small form-factor and depth estimation abilities are crucial like endoscopy.

Since the optical masks used in lensless cameras do not have a focusing action similar to that of a lens, the resultant measurement obtained from a lensless camera is a multiplexed representation of the scene that does not bear any resemblance to the photograph of the same scene. In order to recover the scene, we need to algorithmically demultiplex the measurement. However, designing recovery algorithms even for 2D scenes is challenging, primarily because of the system's poor conditioning, large point spread function (PSF), and extreme multiplexing. Nevertheless, there have been recent works by [6,7,9,10] that have shown that, with strong data-driven priors, one can estimate the 2D scene accurately.

In early works by [2,5,11], it was shown that lensless measurements encode depth information up to a certain depth range. This is due to the fact that the PSFs of lensless imagers scale in size based on the depth of the point source. Thus, a single 2D measurement captured by a lensless camera is actually a compressed representation of the 3D scene. Therefore, to recover the 3D scene from 2D measurement, one needs to solve a highly under-constrained inverse problem, which adds to the challenges imposed on developing lensless recovery algorithms [2,5,11], solving a regularized least squares with strong

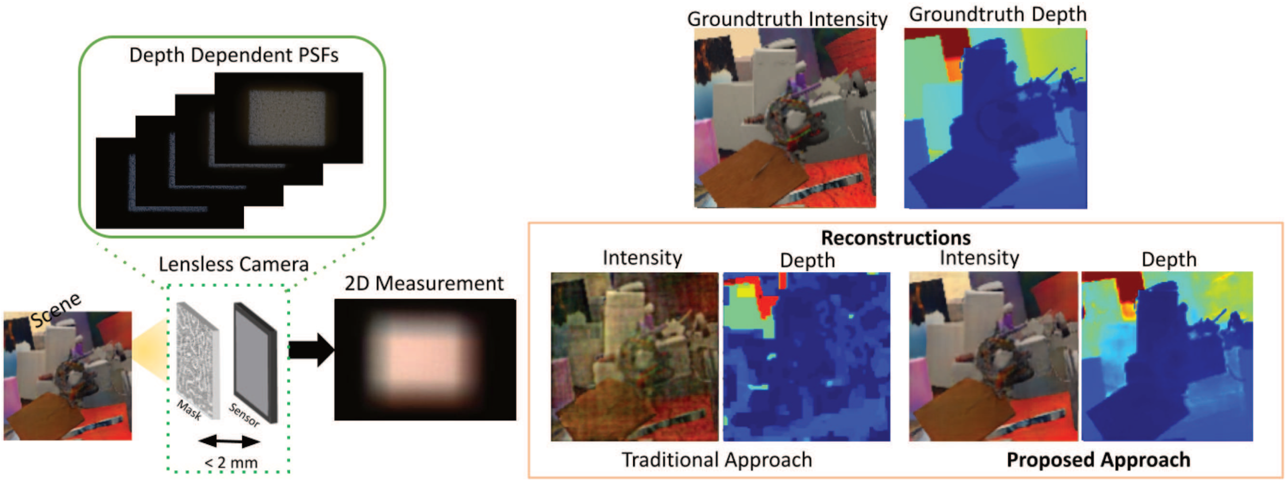

Fig. 1. Lensless 3D imaging. The lensless measurement is a linear combination of the depth dependent point spread function contributed by each scene point. We propose a learning-based approach called FlatNet3D to learn the inverse mapping from a 2D measurement to intensity and depth map.

hand-crafted priors to recover the 3D volume from the 2D measurements. However, these methods mainly work for sparse scenes and are extremely slow due to their iterative nature [12], and they use an optimization-based approach to solve for both intensity and absolute depth from a single lensless measurement. However, they only showed results for separable lensless models, which are known to have poor system characteristics [5]. References [13,14] used a programmable mask to improve the conditioning of the lensless system for depth recovery. But their system has a much thicker form-factor and requires multiple captures to estimate an accurate depth map.

The above challenges make the development of a robust 3D scene estimation algorithm for single-shot lensless imaging the need of the hour. Keeping this in mind, we propose an end-to-end trainable deep network that learns direct mapping from a single lensless measurement to the scene intensity and the absolute depth map. We call this deep network *FlatNet3D*, which employs an efficient physics-based 3D mapping stage followed by a fully convolutional network, which is trained in a coupled fashion. Finally, to verify the robustness and efficiency of FlatNet3D, we perform extensive experiments on challenging simulated and real scenes captured using PhlatCam [5]. Please see Fig. 1. Our contributions are as follows:

- We propose a feed-forward deep network for fast and high-quality intensity and depth reconstruction from single lensless capture.
- At the core of our approach is the proposed learned physics-based 3D mapping stage that brings the measurement to an intermediate 3D volume. It is extremely efficient, requires very few parameters to learn, and allows fast implementation in the Fourier domain.
- We perform our evaluations on both simulated and real captures under various scenarios.
- Finally, we show relevant applications that can benefit from the 3D imaging abilities of the lensless system.

A. Related Works

1. *Mask-Based Lensless Imaging*

Ultra-thin mask-based lensless cameras replace the lens of a traditional lens-based system with an optical mask placed close to the sensor. FlatCam [1] used a separable amplitude mask placed approximately a millimeter from the sensor. It was used to show 2D imaging and 3D volume reconstruction in [11]. DiffuserCam [2] used a random off-the-shelf diffuser as a mask placed nearly 10 millimeters from the sensor. The authors demonstrated 3D imaging ability using this prototype. More recently, its ability to do high speed imaging [15] was also demonstrated. Spectral DiffuserCam [16] exploits the multiplexing ability of lensless imagers to do hyper-spectral imaging. PhlatCam [5] was recently proposed and uses a designed phase-mask with specific properties that make solving the inverse problem easier. The authors demonstrated its ability to do both 2D and 3D imaging.

2. *Learning for Lensless Imaging*

Recently numerous learning-based algorithms have been proposed for lensless scene reconstructions. Reference [9] proposed a feed-forward deep network that performed photorealistic 2D scene reconstructions from separable mask FlatCam measurements. Reference [7] proposed an unrolled deep network that performed 2D intensity reconstructions from DiffuserCam measurements. Recently, [6] proposed FlatNet that was shown to perform 2D intensity reconstructions for a general lensless system. Reference [10] proposed an unsupervised iterative approach that exploits deep image prior for lensless reconstructions. Although the above deep-learning-based techniques are for 2D scene reconstructions, there are no deep learning based approaches for 3D scene estimation from single-shot lensless captures. Moreover, extending the 2D methods for 3D is not trivial, and a naive extension can lead to significant blow-up in memory requirement and parameter count. In this work, we bridge this gap by proposing a fast photorealistic learning-based

3D scene estimation algorithm. Recently transformers [17] have been shown to be effective for lensless image recognition and 2D scene reconstruction [18,19] that are purely data-driven and do not require modeling the lensless camera and calibrating the PSFs—a condition desirable when accurate calibration of PSFs is not possible. However, these works have not explored 3D lensless imaging. Moreover, transformers used in these works consume a significant amount of memory, and fitting them on a single Titan X GPU, which has been used in this work, is not possible. In contrast, since we have an accurate calibrated PSF, the two-stage approach of FlatNet3D makes the estimation more efficient. First, the efficient 3D mapping stage decouples the depth information from a *single* measurement while revealing the image-like structures. Then, in the second stage, the decoupled volume of image-like features is passed through a U-Net to efficiently estimate RGB and depth map.

3. 3D Lensless Imaging

References [2,5] perform 3D voxel reconstructions from 2D lensless measurement using strong scene priors and traditional optimization routines. Recently, authors in [12] proposed a joint intensity and depth reconstruction framework using alternating optimization framework. However, their approach was only shown for a separable model, and extending it to non-separable model is not trivial. The above works rely on strong hand-crafted priors and traditional optimization routines, and they are iterative in nature. In contrast, our proposed approach is feed-forward, learns the priors from the data, and is extremely fast. Another line of work involves using programmable masks and multiple measurements for 3D estimation from thicker form-factor lensless cameras [13,14]. In [8], the authors use amplitude Fresnel zone aperture (FZA) masks. They use 4–16 radially phase-shifted FZA to combine multiple measurements with a virtual FZA to obtain a refocus stack. Changing the pitch of the virtual FZA post-capture allows them to obtain the refocus stack. However, this method is limited to the FZA pattern and uses multiple measurements. Compared to these works, we show 3D imaging capabilities of lensless cameras for a single passive mask and for an ultra-thin geometry, which is much more challenging.

2. 3D LENSLESS IMAGING—A BRIEF BACKGROUND

Each scene point in the field of view of a lensless camera forms a PSF on the sensor. This PSF is a function of the 3D position of the point. Assuming shift invariance within each scene plane, the measurement recorded at the sensor can be written as

$$Y = \sum_z S(z) * H(z). \tag{1}$$

Here, $S(z)$ represents the scene points at the depth plane located at z, $H(z)$ represents the on-axis PSF corresponding to a point source located at z distance from camera, and the asterisk represents 2D convolution. Y is the 2D measurement recorded at the sensor. The summation over z indicates the super-position of the contribution made by each depth plane. Moreover, if $H_\infty(x, y)$ is the PSF for an on-axis point source at optical infinity, then the PSF for an on-axis point at depth z is given by

$$H(x, y, z) = H_\infty\left(\frac{x}{1+d/z}, \frac{y}{1+d/z}\right), \tag{2}$$

where d is the mask to sensor distance and z is the same depth as before [5]. The above equation indicates that the effect of depth on the PSF is reflected through scaling, and this scaling is more challenging to resolve for ultra-thin lensless cameras. The shift-invariance assumption is valid due to the small thickness and large sensor size of the PhlatCam [5] lensless camera that we have used to demonstrate our experiments.

The objective of this paper is to estimate scene intensity $\sum_z S(z)$ and the depth map $z(i, j)$ corresponding to each scene point. In our experiments, we discretize the depth into planes.

3. FLATNET3D—TWO-STAGE NETWORK FOR LENSLESS 3D IMAGING

FlatNet3D is a two-stage deep network that learns the mapping from a 2D lensless measurement to a scene intensity and depth map. The first stage of the network maps the 2D measurement to a intermediate 3D stack. The second stage learns a mapping from this stack to the intensity and depth map through a fully convolutional network. Finally, the network is trained using losses imposed on the intensity and depth map (see Fig. 2). In this section, we will describe each of these stages in detail.

A. Physics-Based Measurement to 3D Mapping

Lensless measurements, due to global multiplexing, lack local structures [6,7]. Moreover, from Eq. (1), it can be seen that the measurement at the sensor is a compressed representation of the 3D volume. Therefore, we need to map the measurement to an intermediate stage with image-like local structures while simultaneously preserving the 3D information. We do so by solving, for each depth plane, the following approximated regularized 2D least squares:

$$S_E(z) = \underset{S(z)}{\operatorname{argmin}} \, ||Y - H(z) * S(z)||_F^2$$
$$+ \lambda(z) ||P(z) * S(z)||_F^2. \tag{3}$$

Here, the asterisk represents 2D convolution, Y is the lensless measurement, $H(z)$ is the PSF corresponding to depth z, $S(z)$ corresponds to scene points at depth z, and $P(z)$ is a regularization filter. If $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the Fourier transform and its inverse, respectively, then the solution to Eq. (3) can be represented as

$$S_E(z) = \mathcal{F}^{-1}(\mathcal{F}(Y) \odot W(z)), \tag{4}$$

where

$$W(z) = \frac{\mathcal{F}(H(z))^*}{|\mathcal{F}(H(z))|^2 + \lambda(z)|\mathcal{F}(P(z))|^2}. \tag{5}$$

$S_E(z)$ in the above equation resembles a noisy focal stack with scene points at that particular depth z appearing sharp. Once the stack is obtained, we pass it through the next fully convolutional volume processing stage.

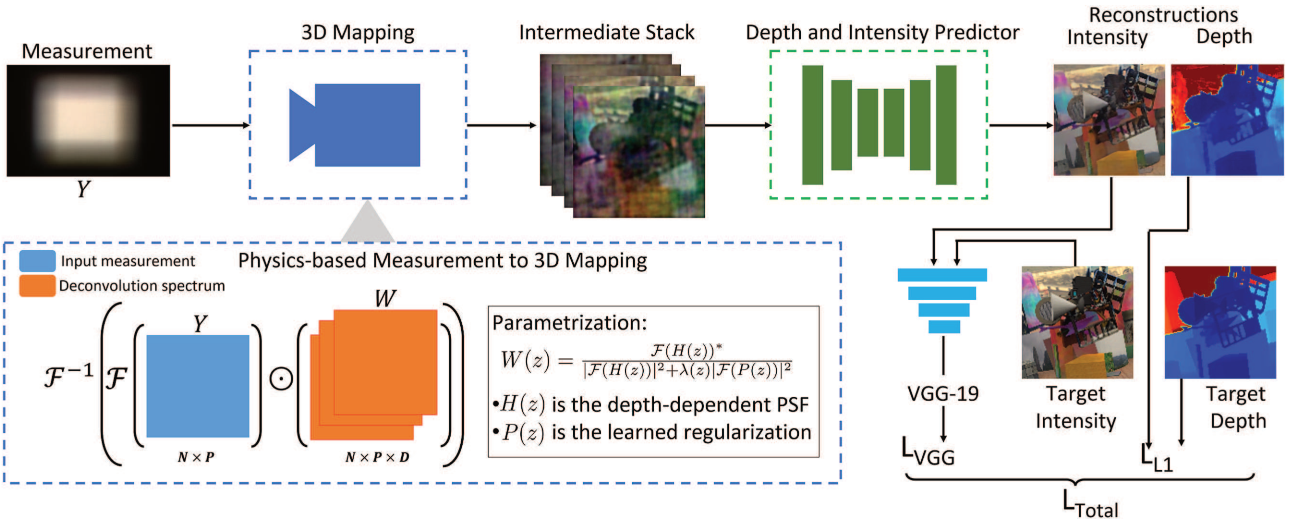

Fig. 2. FlatNet3D. Our proposed network first maps the measurement into an intermediate 3D stack. A convnet then uses this stack to generate intensity and depth estimates. Finally, the entire network is trained in an end-to-end fashion using VGG loss on intensity images and L1 loss on depth maps.

Given that both Eqs. (4) and (5) are fully differentiable, we learn the filters $P(z)$ and the vector $\lambda(z)$. Please note that, unlike the trainable inversion stage of FlatNet [6], the learned mapping corresponding to FlatNet3D has much fewer parameters. This is because of the efficient parameterization of $W(z)$ using Eq. (5), where only $P(z)$ and $\lambda(z)$ are learned, which are typically of much lower dimensions.

B. 3D Stack to Intensity and Depth Prediction

The next stage of FlatNet3D takes in the stack S_E from the previous stage and outputs the depth and intensity estimate. Owing to its large scale success in image-to-image translation, image segmentation, and depth reconstruction problems, we choose a U-Net [20] to learn this mapping from S_E to final intensity and depth map. We have kept the kernel size fixed at 3 × 3, and the number of filters has been gradually increased from 64 to 1024 across five encoder blocks and then reduced back to 64 across 4 decoder blocks. The number of input channels of the U-Net is thrice the total number of discrete depth planes C with each set corresponding to R, G, and B channels of each image in the stack. Since we formulate our depth estimation as a regression problem, the decoder part of the U-Net has four output channels where the first three channels output the RGB intensity estimate and the last one is used to predict the depth of the pixels. Owing to the success of attention gates in previous works, we have used grid-attention proposed by [21] that enables the model to "focus" on the important regions and retain only the necessary activations. Details of the U-Net are provided in Supplement 1.

C. Loss Function

Our model jointly learns to predict the intensity image as well as the depth map. To do so, we use a combination of the following losses.

1. VGG Loss

To learn the intensity mapping, we use the VGG loss proposed in [22], which is known to produce photorealistic images. Let ϕ_j denote the feature map of size $C_j \times W_j \times H_j$ obtained by the jth activation within the VGG19 network. The 22nd convolutional layer in the VGG model has been used as the perceptual output. We then define the VGG loss as the Euclidean distance between the feature representations of a reconstructed image I_{rec} and the reference image I_{ref},

$$\mathcal{L}_{\text{VGG}} = \frac{1}{C_j W_j H_j} ||\phi_j(I_{\text{rec}}) - \phi_j(I_{\text{ref}})||_2^2. \tag{6}$$

2. L1 Loss

We formulate the depth estimation as a regression problem for each pixel. Given the target depth d_{ref} and the predicted depth image d_{pred}, which is also the output of the neural network, the L1 loss is given by the L1 norm between d_{ref} and d_{pred},

$$\mathcal{L}_{L1} = ||d_{\text{ref}} - d_{\text{pred}}||_1. \tag{7}$$

3. Total Loss

The final loss function used for jointly training the network is

$$\mathcal{L}_T = \mathcal{L}_{\text{VGG}} + \alpha \mathcal{L}_{L1}, \tag{8}$$

where α is a scalar hyper-parameter.

4. EXPERIMENTS AND RESULTS

A. Dataset

For all our experiments, we will be focusing on the camera geometry of PhlatCam [5]. PhlatCam is a thin passive mask-based lensless camera. It uses a fixed phase-modulating mask placed very close to the sensor (<2 mm) giving a very thin form-factor. The design of this mask is described in [5] and is based on certain heuristics that allow high-quality imaging. The mask was fabricated using a two-photon lithography 3D printer (Photonic Professional GT, Nanoscribe GmbH). It was printed on a 700 μm thick, 25 mm square fused silica glass substrate using Nanoscribe's IP-DIP photoresist in the Dip-in Liquid Lithography (DiLL) mode with a 63× microscope objective lens. Given that the thickness of PhlatCam is less than 2 mm, we consider a scene depth range of up to 20 cm. Beyond 20 cm, the effect of PSF scaling becomes negligible, and only 2D imaging is possible from single capture. We first capture real PSFs by placing a point source at 25 different depths between 3.6 cm to 20 cm from PhlatCam. Since there is no existing dataset for lensless depth and intensity with measurements and ground truth pairs, we leverage existing RGB-D datasets for generating them. To this end, we use the intensity and disparity images of a subset of the FlyingThings3D [23] dataset to generate lensless measurements using the forward model described in Eq. (1). We use 26,066 RGB-D scenes for this purpose (21,818 for training, 3000 for validation, and 1248 for testing). Given that the FlyingThings3D dataset provides only the disparity, we first clipped the disparities to lie within [10,1000]. This ensures uniformity in the cropped depth images. We calculate the inverse disparity and scale it to the range [0,1]. This is finally followed by a linear map [0,1] to [3.6 cm,20 cm], which is our required depth range. Finally, using the captured PSFs, the scenes, and the depth maps, we simulate the measurements. We add Gaussian noise to the simulated measurement so that the peak signal to noise ratio (PSNR) measurement corresponds to 20–50 dB. We then use this simulated dataset for training and testing purposes. In our experiments, we consider a scene size of 128 × 128.

For testing on real data, we use PhlatCam to capture scenes placed within the above depth range in front of the camera. We provide visual results for these captures later in this section.

B. Training Details

We implemented our models in PyTorch [24]. The Adam optimizer [25] was used to train the network with a learning rate of 10^{-3}. Due to the GPU constraints, we used a batch size of 7. The weight α was varied from 0.001 to 0.002. NVIDIA Titan X GPUs were used for our experiments.

C. Baselines and Metrics

1. Baselines

For comparison with the traditional approach, we used two different methods to obtain a 3D volume or focal stack and used this stack to estimate depth and intensity images. The first approach was to solve, for each depth plane, the Laplacian regularized 2D least squares given by

$$\hat{S}(z) = \underset{S(z)}{\operatorname{argmin}} \; ||Y - H(z) * S(z)||_F^2 + \lambda ||L * S(z)||_F^2, \quad \textbf{(9)}$$

where Y is the measurement, * represents 2D convolution, $S(z)$ are the scene points at depth z, $H(z)$ is the PSF corresponding to depth z, L is a 2D Laplacian filter of size 3 × 3, and λ is a constant. The solution to Eq. (9) is given by

$$\hat{S}(z) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(H(z))^*}{|\mathcal{F}(H(z))|^2 + \lambda|\mathcal{F}(L)|^2} \odot \mathcal{F}(Y)\right). \quad \textbf{(10)}$$

This is the commonly used 2D constrained least squares (CLS) filter used in image processing [26]. The second approach is to use the alternating direction method of multipliers (ADMM) proposed in [2,5]. The ADMM-based approach treats the 3D reconstruction problem as a regularized least squares optimization problem:

$$\hat{S} = \underset{S}{\operatorname{argmin}} \; \frac{1}{2}||Y - \sum_z H(z) * S(z)||_2^2 + \lambda||\Psi(S)||_1 + \lambda_1||S||_1. \quad \textbf{(11)}$$

Here, Ψ is the gradient operator, S is the 3D volume, and $\hat{S}$ is its estimate. The rest of the notations have the same meaning as that in Eq. (9).

The 3D volume obtained by using the above methods may have non-zero values for a pixel at more than one depth plane. Hence, we must apply other methods on top of 3D volume estimation to obtain the intensity image and the depth map. To obtain the same, we have used a graph-cut [27] to minimize the following energy function formulated in [28]:

$$E(x) = \sum_{i \in \mathcal{V}} E_i(x_i) + \lambda \sum_{(i,j) \in \mathcal{E}} E_{ij}(x_i, x_j), \quad \textbf{(12)}$$

where $E(x)$ is the energy of a depth labeling x, x_i is the depth assigned to a pixel $i \in \mathcal{V}$, and $E_i(x_i)$, called the unary potential of a pixel $i \in \mathcal{V}$, is a measure of defocus and is obtained by computing $\exp(-|\nabla I(i)|^2)$ followed by Gaussian averaging over a fixed window. Also, $E_{ij}(x_i, x_j) = |x_i - x_j|$, where $(i, j) \in \mathcal{E}$, the set of edges connecting adjacent pixels. λ is a weighting constant between the unary and pairwise terms. The RGB value of the pixel i in the all-in-focus image is the corresponding RGB value of the x_i th stack.

We call these methods CLS + Graphcut and ADMM + Graphcut depending on the approach used to obtain the 3D volume.

We also make a comparison against a modified version of FlatNet [6], which was originally proposed for 2D scene estimation from lensless measurements. We modify FlatNet by replacing its inversion stage with the learned mapping described in Section 3.A. However, this mapping now uses only the PSF corresponding to the hyperfocal distance. We also modify the perceptual enhancement stage of FlatNet to predict both the RGB and depth map. We call this model FlatNet2D.

2. Metrics

For quantitative evaluation, we use a combination of PSNR (in dB), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) [29] for intensity reconstructions. While PSNR measures the signal distortion, SSIM and LPIPS are useful for quantifying the perceptual quality of the estimates. Higher values of PSNR and SSIM, and lower values of LPIPS, indicate better estimates. For depth estimates, we use root mean square error (RMSE) (in cm) for our evaluation. We report the average metrics evaluated on the simulated test set only. The real data does not have corresponding ground truth. We also report the average inference time for each approach evaluated on a Nvidia Titan X GPU.

D. Comparison with Baselines

1. Quantitative Evaluation

We present the quantitative comparison of our approach against the baselines. Table 1 reports the average metrics evaluated on the simulated test set described in Section 4.A. We can see that FlatNet3D clearly outperforms traditional optimization-based approaches like ADMM + Graphcut or CLS + Graphcut in terms of accuracy and speed. This is primarily because both ADMM and CLS lead to noisy 3D volume predictions, and as a result, the subsequent Graphcut-based approach is unable to extract meaningful depth and intensity from this volume.

Table 1. Quantitative Comparison with Other Approaches[a]

Method	PSNR (dB)	SSIM	LPIPS	RMSE (cm)	Inference Time (s)
CLS + Graphcut	16.24	0.56	0.51	4.87	1.21
ADMM + Graphcut	15.94	0.63	0.32	5.26	10.26
FlatNet2D	19.86	0.65	0.30	1.92	**0.011**
FlatNet3D	**21.91**	**0.79**	**0.14**	**1.42**	0.013

[a] We report the average metrics for the proposed FlatNet3D along with that of the baselines evaluated on the simulated test set. FlatNet3D outperforms all the other baselines in terms of model accuracy.

Among the learning-based approaches, FlatNet3D outperforms FlatNet2D because the latter is not able to extract the depth information accurately from one PSF. Moreover, FlatNet3D

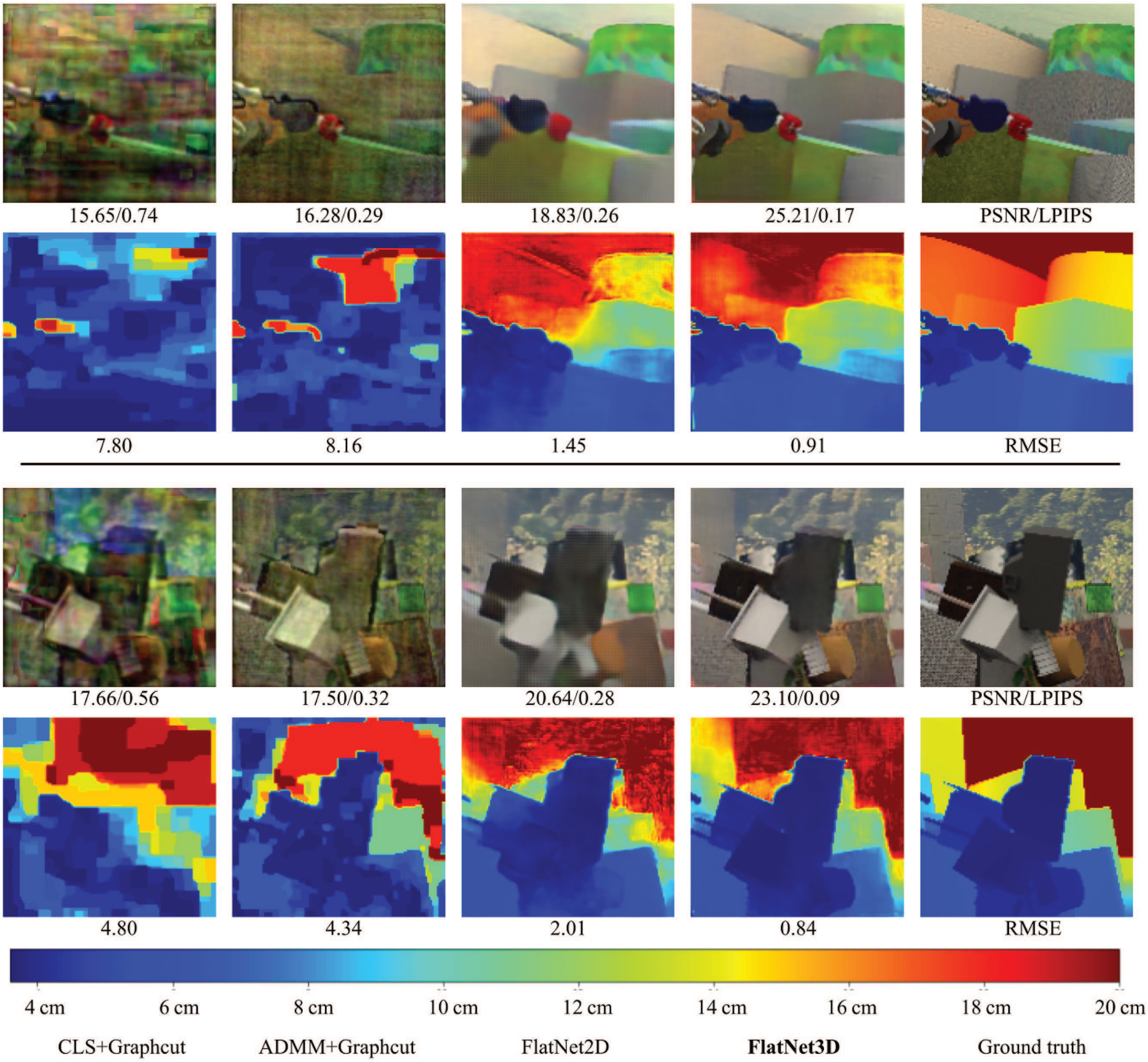

Fig. 3. Comparison on simulated dataset. FlatNet3D provides more photorealistic intensity images and accurate depth maps.

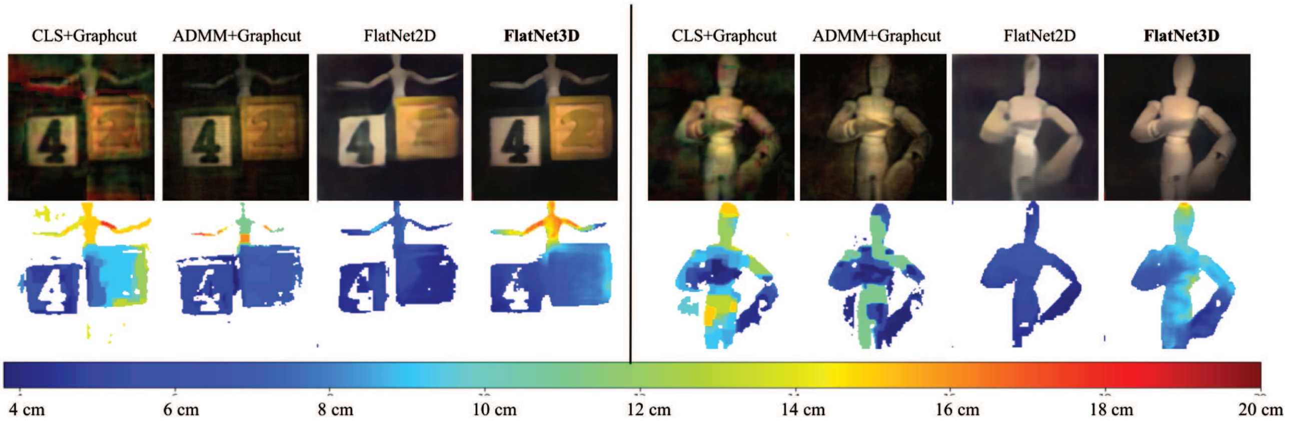

Fig. 4. Comparison on real captures. We show real results for two scenes. FlatNet3D provides better contrast for intensity images and cleaner depth maps for both the scenes.

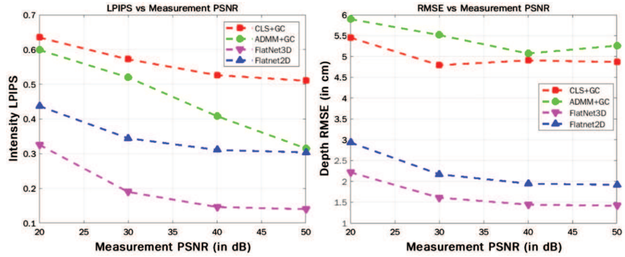

Fig. 5. Noise analysis. We vary the measurement noise and evaluate the performance of all the methods. FlatNet3D outperforms all other methods.

is nearly 100x faster than ADMM + Graphcut and nearly 10x faster than CLS + Graphcut.

2. *Qualitative Evaluation*

In this subsection, we provide visual results for the proposed FlatNet3D against the baselines. In Fig. 3, we show the intensity and depth map reconstructions for two simulated scenes. FlatNet3D provides better depth maps and perceptual intensity quality as reflected through lower LPIPS and RMSE values. The traditional approaches provide noisier intensity and depth estimates while FlatNet2D is unable to extract sharp image and accurate depth map from a single focal stack image.

We also provide visual comparison on real data captured using PhlatCam [5] in Fig. 4. Similar to results on simulated data, FlatNet3D provides better quality intensity estimates reflected through better contrast. The depth estimates using FlatNet3D have fewer spurious regions especially for the bottom scene that lacks texture. Traditional methods suffer from noise for these data as well, with ADMM + Graphcut performing better than CLS + Graphcut. FlatNet2D is unable to extract sharp intensity image and accurate depth map from a single focal stack image. Since dark regions can have ambiguous depth values, we throw away the depth values corresponding to dark scene pixels in Fig. 4. Following [12], we suppress the depth values for pixels with mean intensities (across the color channels) below the image standard deviation.

E. Noise Analysis

We have trained our model with a large range of noise. In this subsection, we evaluate its performance against the baselines for different noise levels. We vary the simulated measurement PSNR from 20–50 dB by varying the noise levels and observe the intensity and depth reconstructions for all the methods. For traditional approaches, we tune the regularization parameters for each noise level separately for optimal performance. In Fig. 5, we show the average LPIPS of intensity estimates and the RMSE of the depth estimate for various noise levels. We only focus on LPIPS since our network parameters are optimized for good perceptual metrics of intensity estimates. Despite using carefully tuned parameters, traditional approaches perform consistently worse than the learning-based approaches both in terms of intensity and depth quality. FlatNet3D outperforms FlatNet2D at all noise levels.

F. Performance on Different Tasks

We evaluate the performance of FlatNet3D on two different RGB-D data based tasks—endoscopy 3D imaging and RGB-D saliency detection.

1. *Endoscopy 3D Imaging*

Getting absolute depth information from endoscopic scenes is of vital importance for diagnosis. In this experiment, we show that low form-factor lensless cameras can allow absolute depth and intensity imaging from endoscopic scenes from single-shot captures. To do this, we simulate lensless measurements from the colon subset of the synthetic EndoSLAM dataset [30] that provides intensity images along with relative depth maps. We use 5000 colon samples for training and 100 samples each from colon, small intestine, and stomach scenes for testing. We first undistort the input images using the calibration files provided by the authors. We then scale the depth to a maximum of 10 cm and fine tune our trained FlatNet3D on this dataset. Figure 6 shows visual results for the intensity and depth maps for various frames. Despite the scenes being extremely low in texture, FlatNet3D is able to extract meaningful absolute depth information from a single measurement.

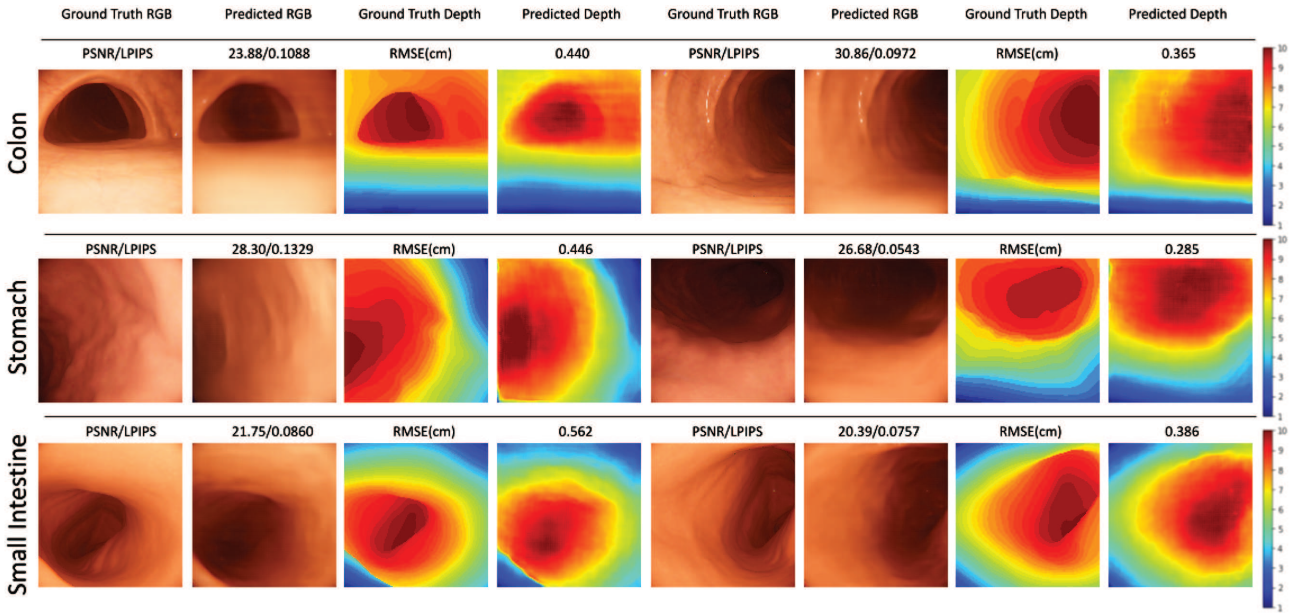

Fig. 6. Performance on EndoSLAM Dataset. We fine tune FlatNet3D on the EndoSLAM RGB-D dataset. It can be seen that FlatNet3D is able to provide a high-quality depth map despite the scenes being low in texture.

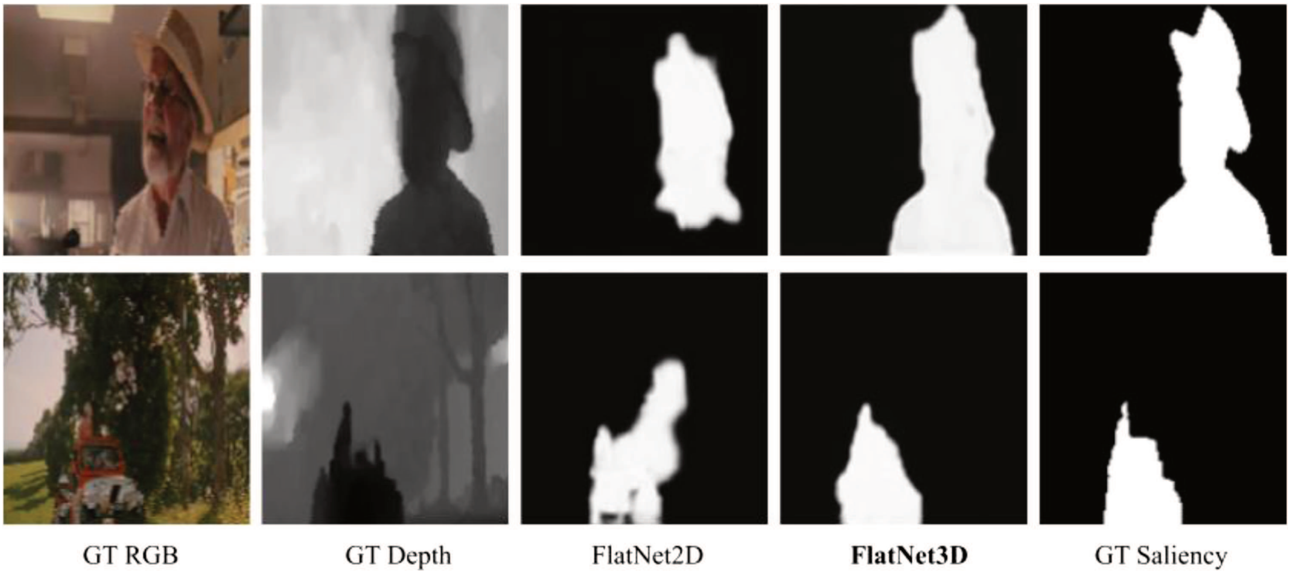

Fig. 7. Performance on salient object detection. Estimates from FlatNet2D and FlatNet3D are used as inputs for saliency detection. Because FlatNet3D predicts both accurate depth map and intensity images, the saliency model trained on its outputs outperforms the saliency model trained on FlatNet2D predictions.

2. *Saliency Detection*

Saliency detection refers to the detection of salient objects/regions within the scene, i.e., the most informative scene regions. Effective detection of salient regions can have numerous applications like object detection, image compression, and medical diagnosis [31,32]. This task can benefit significantly from 3D information.

Saliency detection is a task that can benefit from 3D information. Therefore, we use the predicted intensity and depth map from FlatNet3D for salient object detection. We use the RGB-D saliency dataset proposed in [33]. We first simulate lensless measurements from these RGB-D images and then apply FlatNet3D to them to predict the intensity and depth map. We then use these intensity and depth maps to learn the saliency map using U-Net. In Fig. 7, we compare the performance of saliency detection. As can be observed, saliency detection from the predictions of FlatNet3D is better than the saliency detection from the predictions of FlatNet2D. This indicates that the better depth and intensity predictions from FlatNet3D are crucial for high-quality saliency detection. Although we have shown FlatNet3D-based saliency detection for natural scenes, it can be extended for endoscopic scenes as well given enough labeled endoscopic samples. Such a saliency detector would be useful in assisting physicians in diagnosis, and designing it would be an interesting future research direction.

5. DISCUSSION AND CONCLUSION

In this work, we proposed a novel deep network for intensity and depth estimation from monocular lensless captures. Our method exploits the scaling of lensless PSF at close depth ranges for this estimation. Unlike traditional methods like [2,5,12], our method exploits the prior in the data for doing so, leading to superior quality intensity and depth estimates. The key component of our approach is the physics-based learned 3D stack mapping stage, which is very efficiently parameterized. FlatNet3D is a step toward making ultra-thin light-weight 3D imaging systems more ubiquitous.

FlatNet3D's performance is limited by the lensless camera's physical geometry constraints—beyond a certain depth, PSF scale becomes insensitive to depth. This implies that the performance of depth estimation worsens with an increase in depth values. However, absolute depth estimation for small ranges can be very useful for systems like endoscopes where form-factor is a serious constraint and the scenes are typically very close to the camera. For such short ranges, due to the PSF scaling cue, getting depth from a single-shot lensless capture is much better posed than getting the same from single-shot lens-based captures. In future, it would be interesting to explore depth estimates using multiple lensless cameras, monocular cues, and dynamic mask patterns similar to [13].

Funding. National Institutes of Health (R21EY029459); National Science Foundation (IIS-1652633, IIS-1730574); Defense Advanced Research Projects Agency (HR0011-17-C0026); Qualcomm Technologies, Inc (Qualcomm Innovation Fellowship 2020 India); IITM Pravartak Technologies Foundation.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

[†]These authors contributed equally to this work.

REFERENCES

1. M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "FlatCam: thin, lensless cameras using coded aperture and computation," IEEE Trans. Comput. Imaging **3**, 384–397 (2017).
2. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "DiffuserCam: lensless single-exposure 3D imaging," Optica **5**, 1–9 (2018).
3. V. Boominathan, J. K. Adams, M. S. Asif, B. W. Avants, J. T. Robinson, R. G. Baraniuk, A. C. Sankaranarayanan, and A. Veeraraghavan, "Lensless imaging: a computational renaissance," IEEE Signal Process Mag. **33**, 23–35 (2016).
4. E. J. Tremblay, R. A. Stack, R. L. Morrison, and J. E. Ford, "Ultrathin cameras using annular folded optics," Appl. Opt. **46**, 463–471 (2007).
5. V. Boominathan, J. Adams, J. Robinson, and A. Veeraraghavan, "PhlatCam: designed phase-mask based thin lensless camera," IEEE Trans. Pattern Anal. Mach. Intell. **42**, 1618–1629 (2020).
6. S. S. Khan, V. Sundar, V. Boominathan, A. Veeraraghavan, and K. Mitra, "FlatNet: towards photorealistic scene reconstruction from lensless measurements," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 1934–1948 (2022).
7. K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," Opt. Express **27**, 28075–28090 (2019).
8. T. Shimano, Y. Nakamura, K. Tajima, M. Sao, and T. Hoshizawa, "Lensless light-field imaging with Fresnel zone aperture: quasi-coherent coding," Appl. Opt. **57**, 2841–2850 (2018).
9. S. S. Khan, V. Adarsh, V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards photorealistic reconstruction of highly multiplexed lensless images," in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 7860–7869.
10. K. Monakhova, V. Tran, G. Kuo, and L. Waller, "Untrained networks for compressive lensless photography," Opt. Express **29**, 20913–20929 (2021).
11. J. K. Adams, V. Boominathan, B. W. Avants, D. G. Vercosa, F. Ye, R. G. Baraniuk, J. T. Robinson, and A. Veeraraghavan, "Single-frame 3D fluorescence microscopy with ultraminiature lensless flatscope," Sci. Adv. **3**, e1701548 (2017).
12. Y. Zheng and M. S. Asif, "Joint image and depth estimation with mask-based lensless cameras," IEEE Trans. Comput. Imaging **6**, 1167–1178 (2020).
13. Y. Hua, S. Nakamura, M. S. Asif, and A. C. Sankaranarayanan, "Sweepcam-depth-aware lensless imaging using programmable masks," IEEE Trans. Pattern Anal. Mach. Intell. **42**, 1606–1617 (2020).
14. Y. Zheng, Y. Hua, A. C. Sankaranarayanan, and M. S. Asif, "A simple framework for 3D lensless imaging with programmable masks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2603–2612.
15. N. Antipa, P. Oare, E. Bostan, R. Ng, and L. Waller, "Video from stills: lensless imaging with rolling shutter," in *IEEE International Conference on Computational Photography (ICCP)* (IEEE, 2019), pp. 1–8.
16. K. Monakhova, K. Yanny, N. Aggarwal, and L. Waller, "Spectral DiffuserCam: lensless snapshot hyperspectral imaging with a spectral filter array," Optica **7**, 1298–1307 (2020).
17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, 2017).
18. X. Pan, X. Chen, T. Nakamura, and M. Yamaguchi, "Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer," Opt. Express **29**, 37962–37978 (2021).
19. X. Pan, X. Chen, S. Takeyama, and M. Yamaguchi, "Image reconstruction with transformer for mask-based lensless imaging," Opt. Lett. **47**, 1843–1846 (2022).
20. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015), pp. 234–241.
21. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: learning to leverage salient regions in medical images," Med. Image Anal. **53**, 197–207 (2019).
22. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision* (Springer, 2016), pp. 694–711.
23. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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* (2016), pp. 4040–4048.
24. A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, 2019), pp. 8024–8035.
25. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).

26. B. R. Hunt, "The application of constrained least squares estimation to image restoration by digital computer," IEEE Trans. Comput. **22**, 805–812 (1973).
27. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1222–1239 (2001).
28. S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3497–3506.
29. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 586–595.
30. K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," Med. Image Anal. **71**, 102058 (2021).
31. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 1597–1604.
32. F. Deeba, S. K. Mohammed, F. M. Bui, and K. A. Wahid, "A saliency-based unsupervised method for angiectasia detection in endoscopic video frames," J. Med. Biol. Eng. **38**, 325–335 (2018).
33. R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *IEEE International Conference on Image Processing (ICIP)* (2014), pp. 1115–1119.