

# **Reconstructing Motion from Rolling shutter Lensless Images and 3D Reconstructions using single shot Lensless Captures**

*A Thesis*

*submitted by*

**DHRUVJYOTI BAGADTHEY**

*in partial fulfilment of the requirements*

*for the award of the degree of*

**BACHELOR AND MASTER OF TECHNOLOGY**



**DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.**

**June 17, 2022**



# THESIS CERTIFICATE

This is to certify that the thesis titled , submitted by **Dhruvjyoti Bagadthey**, to the Indian Institute of Technology, Madras, for the award of the degree of **Dual Degree (Bachelor of Technology + Master of Technology)**, is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Dr. Kaushik Mitra**  
Research Guide  
Assistant Professor  
Dept. of Electrical Engineering  
IIT-Madras, 600036  
Place: Chennai  
Date: June 17, 2022



## ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude and appreciation to my advisor Dr. Kaushik Mitra for his guidance throughout my research work. I am thoroughly indebted to him for the support, motivation, and words of encouragement throughout the ups and downs of my project period. I would like to thank him for providing a nurturing atmosphere and unique opportunities for my growth. I have learned a lot from him in my professional journey.

I would especially like to thank Salman Siddique Khan from the Computational Imaging Group for his mentorship, brainstorming sessions, and direction to proceed and give a proper shape to this work. His insights, his knowledge, and his intuition have facilitated me to develop a temperament to approach research problems. His patience, perseverance, and never-say-never approach to problems have helped me to grow as an individual. This work would not be possible without his help.

I would like to thank Sanjana S Prabhu, my collaborator in the 3D Lens-less project, my lab-mate, and my friend with whom we advanced the 3D Lens-less project. I would like to mention her cooperation and the numerous (many times last-minute) experiments that we would conduct before a deadline. I would also like to thank her for the times she filled in for me when I was unavailable.

I would also like to acknowledge the support of our collaborators at Rice University: Dr. Vivek Boominathan and Prof. Ashok Veeraraghavan without whom this project would not be possible.

I would like to thank IIT Madras for providing a holistic environment for personal and professional growth. I would like to mention D Tony Fredrick, Manogna K and Sarah, my lab-mates, and my friends for their support, encouragement, and companionship that were instrumental in making my life at IIT Madras a memorable journey.

Lastly, I would like to acknowledge the support of my parents who have always been the pillars of strength in my life.



# ABSTRACT

**KEYWORDS:** Lensless Imaging, High speed video reconstruction, High speed motion classification, 3D Lensless Imaging, Fourier inversion, Computational Photography, Point Spread Function

Lensless Imaging has recently gained momentum since it has emerged as an attractive solution for ultra-compact inexpensive imaging. Since the lensless measurements have global multiplexing, lensless cameras can be leveraged for compressive video sensing to achieve a higher spatio-temporal resolution than possible with a lens-based camera. A high frame-rate video can be encoded in a single capture using a rolling shutter CMOS sensor. However the resulting reconstruction problem is highly ill-posed and traditional methods often require hand crafted image priors and highly sparse scenes and long iterations resulting in higher inference time. In this work we propose an end-to-end Deep network for video reconstruction using single rolling shutter lensless captures. Our approach provides 10x frame rate increase, better perceptual quality than traditional methods in a much lesser inference time. We compare our approach for two traditional methods and evaluate its performance on a simulated sparse dataset using Phlatcam.

Hightspeed motion recognition is imperative for various applications like AR/VR, combat elements, motion tracking devices etc requiring high bandwidth where traditional camera fails to capture high speed motion. Since lensless cameras provide global multiplexing, we can leverage the compressive sensing capabilities of rolling shutter lensless cameras for this purpose effectively. In this work, we propose an end-to-end trained CNN based network that can classify activity from a single rolling shutter lensless capture. We also compare our model with the best case classifier, i.e a traditional lens-based camera having higher bandwidth.

Lensless cameras encode depth information in their measurements for a certain depth range. Previous works have shown that this encoded depth can be used to perform 3D reconstruction of close-range scenes. However, these traditional approaches for 3D reconstructions are typically iterative and optimization based that 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 an end-to-end trainable feed-forward deep network *FlatNet3D* - that can estimate both depth and intensity directly from a single lensless capture. Our algorithm is fast, efficient and we demonstrate the high-quality results validated using both simulated and real scenes captured by PhlatCam. We also show the effectiveness of our model on one of the important application of our framework, medical endoscopy.



# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>i</b>
<b>ABSTRACT</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>vii</b>
<b>LIST OF FIGURES</b>	<b>x</b>
<b>ABBREVIATIONS</b>	<b>xi</b>
<b>NOTATION</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background and Problem Statement</b>	<b>5</b>
2.1 The Point Spread Function . . . . .	5
2.2 The 2D Forward Model . . . . .	6
2.3 The Rolling Shutter Forward Model . . . . .	7
2.4 The 3D Forward Model . . . . .	10
2.5 Summary of Assumptions . . . . .	10
<b>3 Related work</b>	<b>13</b>
3.1 Mask Based Lensless Imaging . . . . .	13
3.2 Learning for Lensless Imaging . . . . .	13
3.3 High Speed Lensless Imaging . . . . .	14
3.4 Motion Recognition . . . . .	14
3.5 3D Lensless Imaging . . . . .	15
<b>4 Proposed Method</b>	<b>17</b>
4.1 Highspeed video reconstruction from single rolling shutter lensless captures . . . . .	17
4.1.1 MeasNet . . . . .	17

4.1.2	Trainable Inversion . . . . .	18
4.1.3	EnhanceNet . . . . .	18
4.1.4	Loss Function . . . . .	18
4.2	Motion recognition from single rolling shutter lensless captures . .	19
4.2.1	ClassificationNet . . . . .	19
4.2.2	Loss Function . . . . .	20
4.3	3D Reconstruction from single lensless measurements . . . . .	20
4.3.1	Physics-based measurement to 3D mapping . . . . .	21
4.3.2	3D stack to intensity and depth prediction . . . . .	21
4.3.3	Loss Function . . . . .	22
<b>5</b>	<b>Experiments and Results</b>	<b>23</b>
5.1	Implementation Details . . . . .	23
5.2	Baselines . . . . .	25
5.2.1	Highspeed Lensless reconstruction . . . . .	25
5.2.2	Highspeed Activity recognition . . . . .	25
5.2.3	3D Lensless reconstruction . . . . .	25
5.3	Metrics: . . . . .	27
5.4	Results . . . . .	27
5.4.1	Comparison with baselines . . . . .	27
5.4.2	Ablation experiments . . . . .	31
5.4.3	Applications . . . . .	32
<b>6</b>	<b>Conclusion and Future work</b>	<b>35</b>

## LIST OF TABLES

5.1	<b>Comparison with the best case classifier</b> The accuracies here are calculated as $100 \cdot (1 - \text{top-k error})$ . . . . .	29
5.2	<b>Quantitative comparison with other approaches.</b> A comparison of the average metrics for the proposed FlatNet3D along with the baselines evaluated on the simulated test set. . . . .	30



# LIST OF FIGURES

1.1	<b>Lensless Camera vs Lens Based cameras</b> Lensless cameras greatly facilitate miniaturization of cameras since the form factor is no longer limited by a lens. This makes low cost fabrication possible and so the camera can even be made physically flexible. . . . .	1
1.2	<b>Lensless Camera Applications</b> Mask based Lensless cameras find applications in IoT devices, medical endoscopy, AR/VR etc where form factor is a major constraint. . . . .	2
2.1	<b>PSF variations of Phlatcam</b> . . . . .	5
2.2	<b>2D Lensless measurements:</b> As we see, the measurement bears no resemblance to the actual scene and lacks local features due to which, we cannot use learning based methods directly. As we shall see later, this motivates the use of a Fourier based inversion stage. . . . .	6
2.3	<b>The Camera Rolling Shutter</b> Temporal variation of a typical rolling shutter. Orange denotes a read operation, grey denotes currently exposing. . . . .	8
2.4	<b>Rolling Shutter Lensless Visualisation: Vertical motion</b> Here we show an example of the rolling shutter lensless pipeline using Phlatcam and a simple scene. the scene irradiance is zero except in the 20th,50th and 70th (coloured correspondingly) time sample to represent motion in the direction of the shutter. We can clearly discern the order in which the blobs appear exactly by using the rolling shutter measurement since the pixel values are related to the temporal variation of the shutter but we cannot do the same for the global shutter measurement. . . . .	8
2.5	<b>Rolling Shutter Lensless Visualisation: Horizontal motion</b> Here we show an example of the rolling shutter lensless pipeline using Phlatcam and a differnet scene. the scene irradiance is zero except in the 20th and 50th (coloured correspondingly) time sample and the motion is in a perpendicular direction to the shutter. . . . .	9
2.6	<b>3D Lensless PSFs</b> An extremely simplified scene for Visualization. The PSF pattern at 1cm(magenta) is a scaled up version of the PSF pattern at 20cm(yellow). This difference helps us distinguish the depth of the two points. Actual scenes are much more complex and finding depth becomes non trivial. . . . .	10
4.1	<b>Highspeed Reconstruction framework</b> Our proposed network first maps the measurement slivers into an intermediate measurement stack. The measurement stack is then deconvolved and passed through a convnet for perceptual enhancement of video. . . . .	17

4.2	<b>FlatNet3D.</b> 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. . . . .	20
5.1	<b>Qualitative comparison on simulated Youtube VOS images.</b> Our method surpasses both the baselines in terms of perceptual quality and temporal consistency. . . . .	28
5.2	<b>Qualitative comparison on simulated Youtube VOS images.</b> Our method surpasses both the baselines in terms of perceptual quality and temporal consistency. . . . .	29
5.3	<b>Visual comparison on simulated dataset</b> . . . . .	30
5.4	<b>Qualitative comparison on real captures.</b> We show real result for two scenes. FlatNet3D provides better contrast for intensity images and cleaner depth maps for both scenes. . . . .	31
5.5	<b>Noise ablation.</b> We vary the measurement noise and evaluate the performance of all the methods. . . . .	32
5.6	<b>Performance on EndoSLAM Dataset.</b> We have finetuned FlatNet3D on the EndoSLAM RGB-D dataset. It can be seen that FlatNet3D is able to provide high quality depth maps despite the scenes being extremely low in texture. . . . .	33

## ABBREVIATIONS

<b>PSF</b>	Point Spread Function
<b>AR</b>	Augmented Reality
<b>VR</b>	Virtual Reality
<b>PCB</b>	Printed Circuit Board
<b>CMOS</b>	Complementary Metal-Oxide Semiconductor
<b>RGB-D</b>	Red Green Blue - Depth
<b>FoV</b>	Field of View
<b>FISTA</b>	Fast Iterative Shrinkage-Thresholding Algorithm
<b>ADMM</b>	Alternating Direction Method of Multipliers
<b>FFT</b>	Fast Fourier Transform
<b>DFT</b>	Discrete Fourier Transform
<b>LPIPS</b>	Learned Perceptual Image Patch Similarity
<b>SSIM</b>	Structural Similarity Index
<b>RMSE</b>	Root Mean Square Error
<b>PSNR</b>	Peak Signal to Noise Ratio
<b>GPU</b>	Graphics Processing Unit





## NOTATION

$\mathcal{F}$	Discrete Fourier Transform
$\mathcal{F}^{-1}$	Inverse Discrete Fourier Transform
$\phi$	Multiplexing matrix, a general linear transformation
$C(.)$	Crop operator
$*$	Convolution operation
$\odot$	Hadamard product
$H$	Point Spread Function
$z$	dept value
$Y$	2D lensless measurement
$X$	2D scene
$V$	3D scene volume
$N$	additive noise
$S$	Shutter mask
$y$	flattened and vectorised lensless measurement

# CHAPTER 1

## Introduction

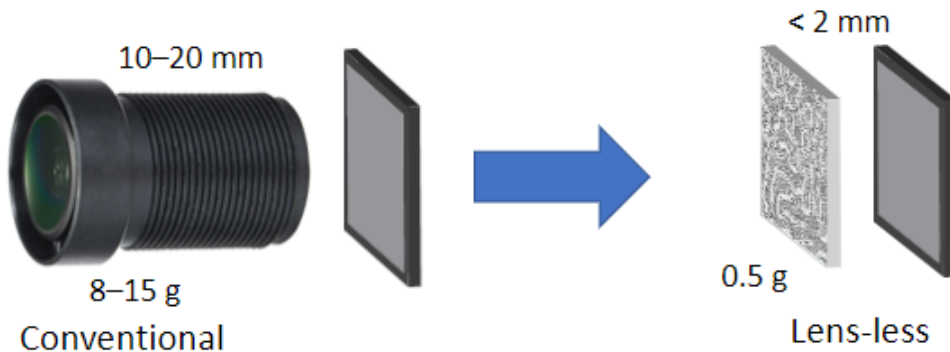


Figure 1.1: **Lensless Camera vs Lens Based cameras** Lensless cameras greatly facilitate miniaturization of cameras since the form factor is no longer limited by a lens. This makes low cost fabrication possible and so the camera can even be made physically flexible.

It is a well known fact that the lens accounts for more than 90% of the camera's weight, volume and cost. Even though there have been advancements in miniaturising lenses and the overall camera design, the fundamental laws of Physics (like the diffraction laws) impose an upper limit to the size, resolution, form factor etc that hinders further miniaturisation. This physical limit makes the lens-based cameras unsuitable for applications where form factor, size and cost are the primary constraints such as micro robotics, pill-endoscopy, mini-drones, AR/VR, wearables.

Over the last decade, lensless imaging systems have emerged as a possible solution for ultra thin, light weight and cost effective imaging. The basic idea common to all lensless frameworks is to replace the focusing element, the lens with a multiplexing element, typically an optical mask that provides cues for reconstructing the original scene using a computational algorithm. Moreover, lensless design permits an inexpensive fabrication method compatible with the conventional pcb based semiconductor fabrication technology allowing it to exploit its scaling, compatibility, cost advantages and flexibility.

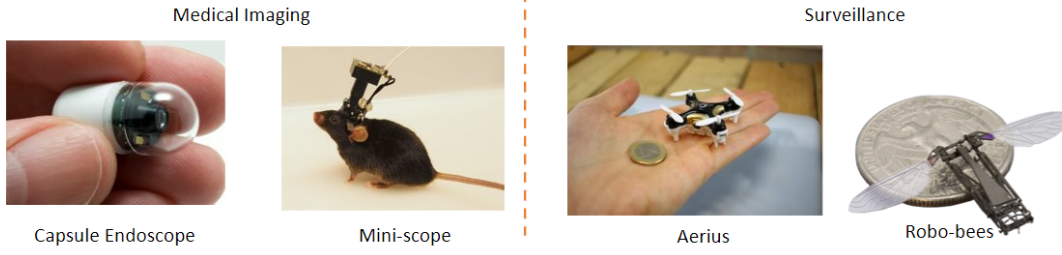


Figure 1.2: **Lensless Camera Applications** Mask based Lensless cameras find applications in IoT devices, medical endoscopy, AR/VR etc where form factor is a major constraint.

Traditional high-speed video cameras are difficult and expensive to fabricate due to the limited camera bandwidth. Most CMOS optical sensors are low-cost, easy to fabricate and have a rolling shutter measurement system. These rolling shutter measurements encode some motion information however it fails to capture motion in other directions as well as the presence of objects depending on their trajectories with respect to the camera. This makes recovering the original motion impossible. This problem can be circumvented by using global multiplexing of scene points in the measurement. In this work, we refer rolling shutter lensless measurement as a single measurement and a rolling shutter lensless video as the video of measurements that is collected from a low fps lensless rolling shutter camera.

Due to absence of a focusing element, the lensless measurements have high global multiplexing which means that a pixel can contain information from many scene points resulting in a measurement that bears no resemblance to the original scene. Designing recovery algorithms, even for 2D scenes is difficult, primarily because of the system's poor conditioning, large PSF and extreme multiplexing although recent works by [1, 2, 3, 4] have shown that one can estimate the 2D scene accurately using strong data-driven priors. This global multiplexing provides an excellent opportunity for reconstructing high-speed videos from low frame-rate rolling shutter lensless videos.

Depth estimation from images is a classic vision problem that has its applications in computer aided diagnosis, robotics, and autonomous systems with many of these systems having strict form-factor and weight constraints. Using existing depth estimation methods like time-of-flight sensors, structured-light or stereo cameras are not feasible, 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 becomes essential. It has been shown in [5, 6, 7] that the lensless measurements encode depth information within a certain depth range. This is because the PSFs scale in size with the depth of the point source. Thus a single lensless measurement is actually a compressed and multiplexed representation of a 3D scene. Therefore, to reconstruct a 3d scene, one needs to solve a highly underconstrained inverse problem on top of the challenges already described for 2D scene reconstructions. [5, 6, 7] solve it using slow iterative methods and strong hand crafted priors which work only for highly sparse scenes and fail to generalise otherwise. [8] uses an alternating optimization approach to solve for both intensity and absolute depth from a single lensless measurement. However, they have only shown results for lensless models with seperable masks which are known to have poor system characteristics.

Keeping this in mind, We propose an end-to-end feed forward convolutional neural network for single shot reconstruction of the all-in-focus image and the depth map from a single Lensless measurement. We refer this network as *FlatNet3D* which combines an efficient implementation of a Fourier based inversion stage followed by a robust fully convolutional network for reconstructing the all in focus image and a depth map. We then evaluate its reconstruction performance over a well known simulated dataset over various noise levels. To verify the robustness of FlatNet3D, we perform extensive experiments over real scenes captured by PhlatCam[7]. Finally to demonstrate an application, we finetune our model over a simulated RGB-D endoscopy dataset.

In summary, this work has the following key contributions:

- To propose an end to end three stage feed forward fully convolutional network for reconstructing rgb frames from a single rolling-shutter lensless measurement with a physics based inversion stage.
- To quantitatively evaluate its performance over simulated images.
- To propose an end to end three stage feed forward fully convolutional network for motion recognition from single rolling-shutter measurements and to compare its performance with a natural image classifier.
- To propose an end to end two stage feed forward fully convolutional network for reconstructing depth map and rgb image from a single lensless measurement with an efficient physics based implementation of the learnable inversion stage.
- To quantatively evaluate its performance on a simulated dataset over various levels of noise.

- To evaluate its performance qualitatively on actual scenes captured by a lensless camera.
- To demonstrate the utility of our model in endoscopy scene reconstructions.

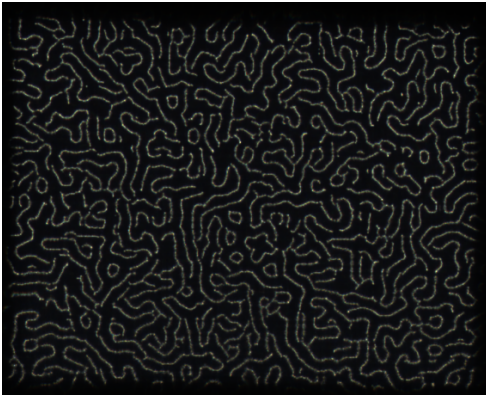
# CHAPTER 2

## Background and Problem Statement

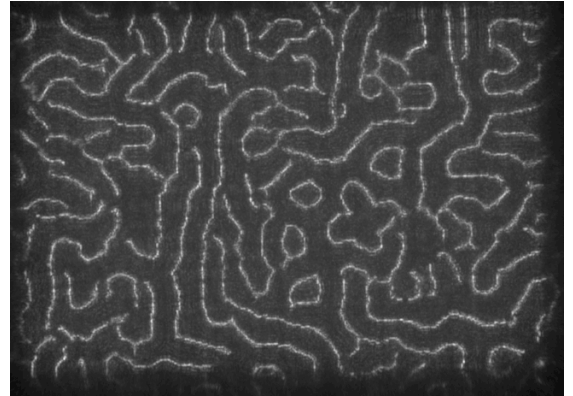
### 2.1 The Point Spread Function

Optical masks can be of various types like an amplitude mask, a phase mask or a diffuser mask based on diffraction. The treatment of those masks is however very similar as far as the camera forward model is concerned. Suppose we place a coherent collimated point light source at a distance  $z$  from the camera's aperture on the camera's axis. The sensor recording reveals a mask pattern formed that we denote as the Point Spread Function (PSF)  $H(z)$ . It can be shown using Fourier optical analysis and the paraxial approximation that the PSF is shift invariant. We also implicitly assume that a scene consists of various incoherent distant scene points which renders the system linear. It should be noted that this linearity is valid only if all the scene points are in the same depth plane. as  $z$  varies, this PSF  $H(z)$  varies with  $z$ , for example as  $z$  decreases, the light source is nearer to the mask and hence the PSF obtained is a scaled up version of  $H(\infty)$ . The depth dependence of the PSF can be denoted as:

$$H_z(x, y) = H_\infty\left(\frac{x}{1 + d/z}, \frac{y}{1 + d/z}\right), \quad (2.1)$$



((a)) Phlatcam A



((b)) Phlatcam B

Figure 2.1: PSF variations of Phlatcam

where  $x, y$  are the  $x$  and  $y$  coordinates of scene points,  $z$  is the depth,  $H_\infty$  is the PSF when the light source is placed at the optical infinity which we will be referring to as the "PSF at infinity" in this work here on.  $d$  is the mask-sensor distance which is typically kept small in case of lensless cameras. We can see that as  $z$  becomes larger and larger, the PSF converges to the PSF at infinity, since the pixel pitch of the camera is finite, this convergence happens at a finite depth  $z_{max}$ . This imposes a physical limitation over the depth based multiplexing of PSFs beyond this  $z_{max}$ . For Phlatcam[7], this  $z_{max}$  is equal to 20cm and hence  $H_{20cm} = H_\infty$ .

## 2.2 The 2D Forward Model

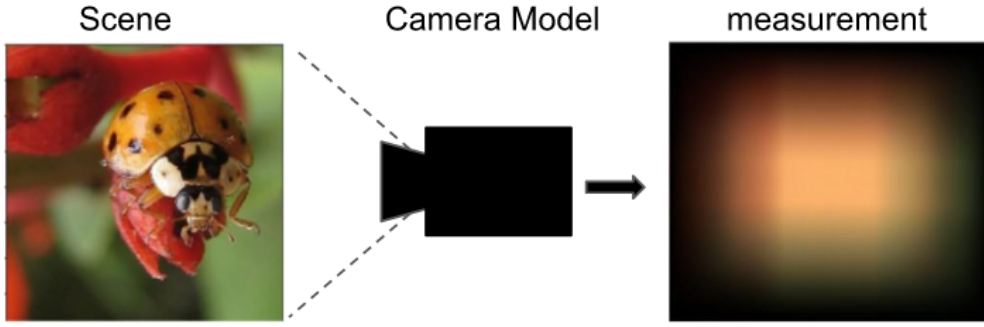


Figure 2.2: **2D Lensless measurements:** As we see, the measurement bears no resemblance to the actual scene and lacks local features due to which, we cannot use learning based methods directly. As we shall see later, this motivates the use of a Fourier based inversion stage.

In this section, we will formulate the camera forward model which is the function that maps a scene to a measurement. The forward model can be mainly categorised as separable and non-separable. Since Phlatcam [7] uses a non-separable mask we shall stick to the non-separable forward models only in this work.

Firstly let us consider a 2D scene  $x$  placed at a depth  $z$ , the camera forward model is a linear system that can be formulated as:

$$y = \Phi_z x + n \quad (2.2)$$

where  $y$  is the measurement,  $n$  is the additive noise, and  $\phi$  is a generalized linear transformation. In general this  $\Phi$  has a large memory footprint and this transformation re-

quires a large computation capacity. Reconstructing a scene with  $\mathcal{O}(N^2)$  scene points requires a  $\Phi$  of the order  $\mathcal{O}(N^4)$ . Due to its computational complexity, the forward model of Phlatcam cannot be simulated easily unlike its separable counterparts.

By adding an aperture over the lensless mask and by exploiting the Linear shift invariant property of PSF it was shown by [7, 6] that the camera forward model can be written as a convolution in the space domain or equivalently an elementwise multiplication in the Fourier domain.

$$Y = H_z * X + N \quad (2.3)$$

Where  $H_z$  is the PSF and  $\phi_z$  is the circulant matrix of  $H_z$ . If the sensor is not large enough compared to the PSF and the FoV, the measurement can overshoot the sensor and the measurement is hence a cropped version of the measurement.

$$Y = C(H_z * X) + N \quad (2.4)$$

where  $C$  is the sensor cropping operation and it causes the system to be non circulant. In Phlatcam however the sensor size is designed to be big enough and hence we will ignore the effects of cropping. This model can be simulated with much lower computation complexity using fourier multiplication and FFT algorithm [9]  $\mathcal{O}(N^2 \log N)$ . Consider a scene placed at a depth greater than or equal to  $z_{max}$  or equivalently the camera's optical infinity. Since we assume that all objects are placed at optical infinity this scene is equivalent to a 2D scene as far as the camera's forward model is concerned since  $H$  for these depths is equal to  $H_\infty$ . This is similar to [10] except that there is no cropping and the PSF is changed.

## 2.3 The Rolling Shutter Forward Model

The above forward model describes a camera with global shutter or a rolling shutter camera with a scene with no motion (or conditions in which a rolling shutter camera works exactly like a global shutter). The camera's shutter function is the temporal variation of the exposure window in the sensor. A typical rolling shutter exposes a row of pixels for  $T_e$  seconds (exposure time) and the next row is exposed after  $T_l$  seconds



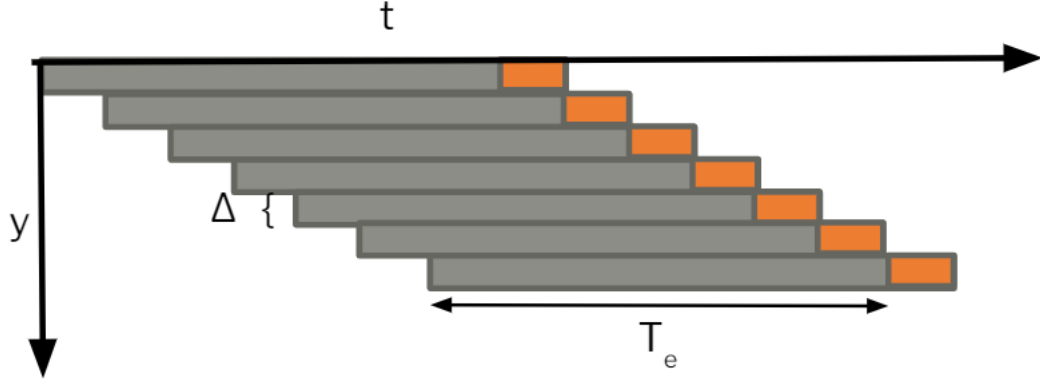


Figure 2.3: **The Camera Rolling Shutter** Temporal variation of a typical rolling shutter. Orange denotes a read operation, grey denotes currently exposing.

(line time). This shutter function can be represented in discrete time using the shutter

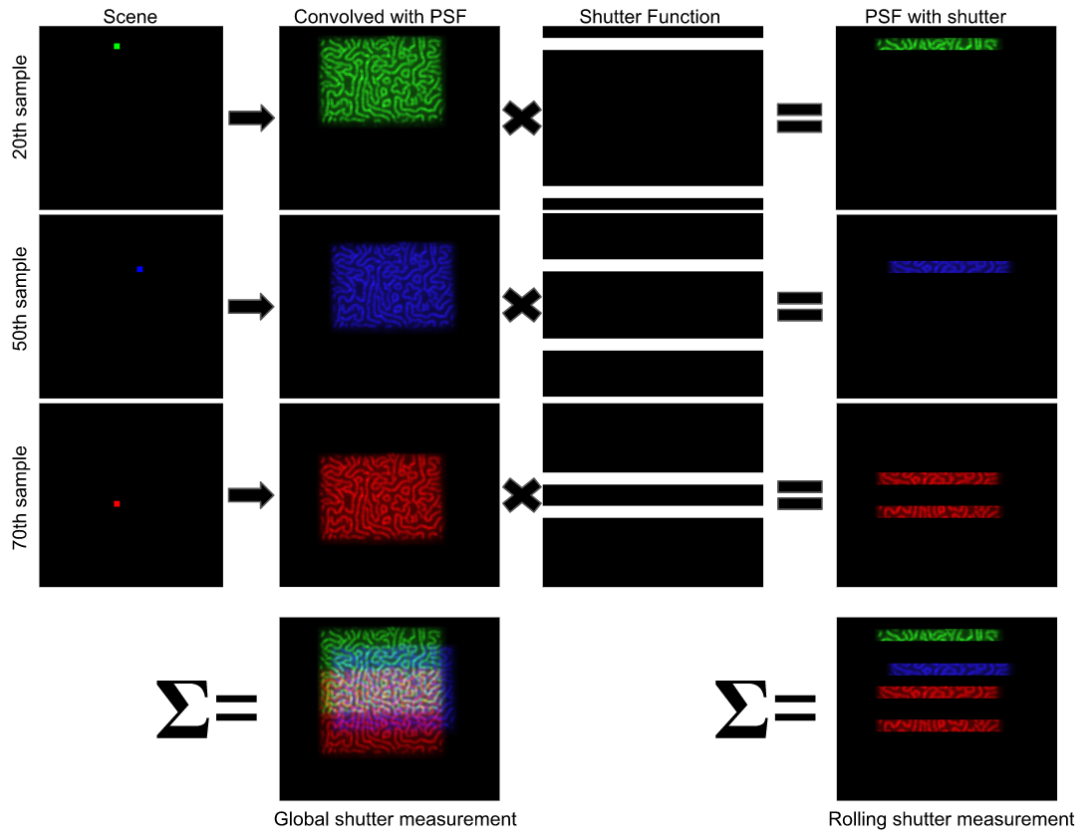


Figure 2.4: **Rolling Shutter Lensless Visualisation: Vertical motion** Here we show an example of the rolling shutter lensless pipeline using Phlatcam and a simple scene. the scene irradiance is zero except in the 20th,50th and 70th (coloured correspondingly) time sample to represent motion in the direction of the shutter. We can clearly discern the order in which the blobs appear exactly by using the rolling shutter measurement since the pixel values are related to the temporal variation of the shutter but we cannot do the same for the global shutter measurement.

mask  $S[n; x, y]$  where  $n$  is the  $n$ th read time or the  $n$ th row read. For most shutter functions,  $S$  is constant along the row hence we can denote it as  $S[n; y]$ . Ignoring the effects of PSF scaling or alternatively considering a scene placed beyond  $z_{max}$  or alternatively considering a 2D scene, the resulting measurement can be formulated as:

$$y = \sum_{n=1}^N S[n; y] \cdot (h(x, y) * v(x, y, n)) + N \quad (2.5)$$

align where  $h$  is the lensless mask,  $v(x, y, n)$  is the scene point at the  $n$ th time sample and  $N$  denotes the number of frames that is a function of  $T_e, T_i$  and the shutter function parameters.

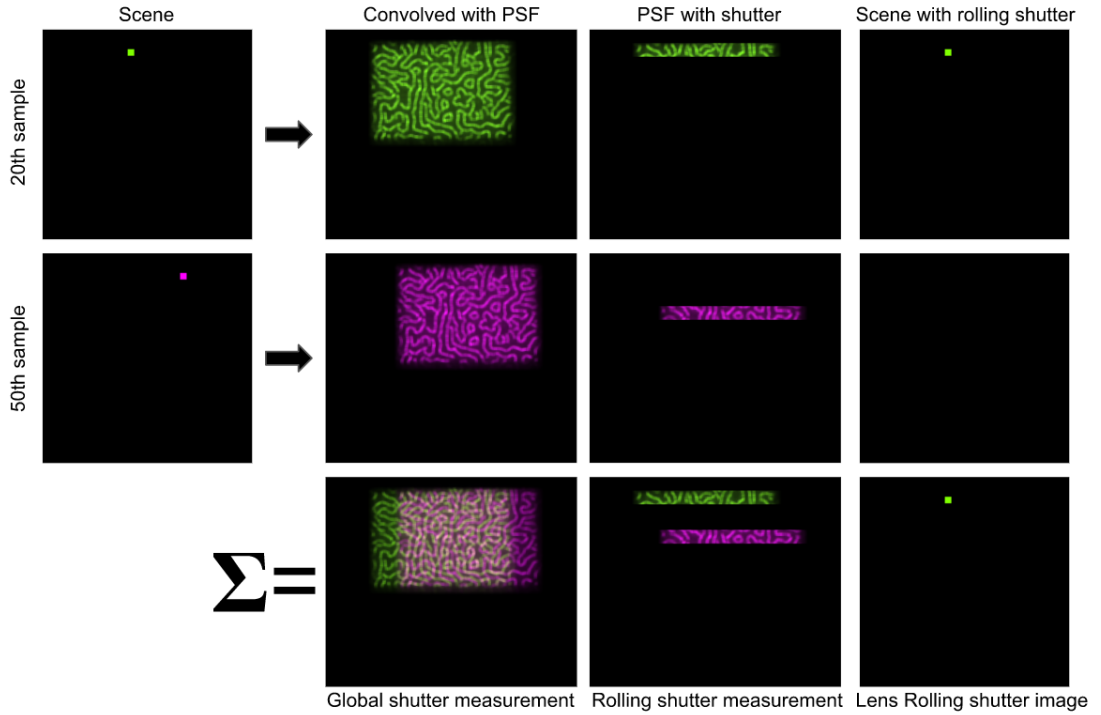


Figure 2.5: **Rolling Shutter Lensless Visualisation: Horizontal motion** Here we show an example of the rolling shutter lensless pipeline using Phlatcam and a different scene. the scene irradiance is zero except in the 20th and 50th (coloured correspondingly) time sample and the motion is in a perpendicular direction to the shutter.

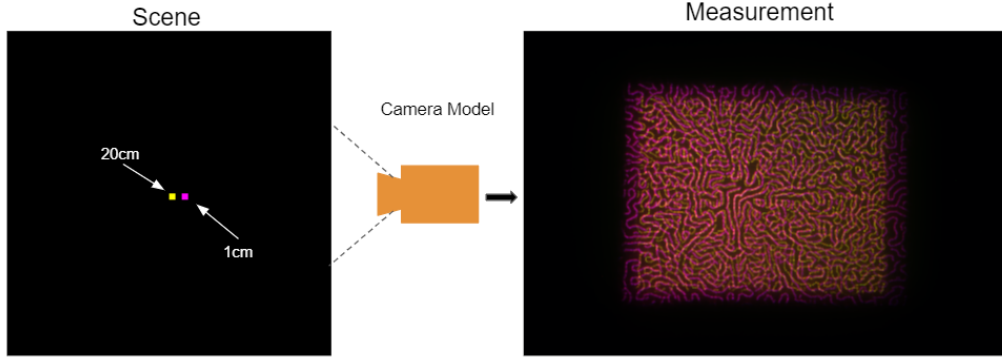


Figure 2.6: **3D Lensless PSFs** An extremely simplified scene for Visualization. The PSF pattern at 1cm(magenta) is a scaled up version of the PSF pattern at 20cm(yellow). This difference helps us distinguish the depth of the two points. Actual scenes are much more complex and finding depth becomes non trivial.

## 2.4 The 3D Forward Model

The PSF scales in the x and y dimensions with depth as 2.1. The measurement can be written as:

$$y = \int_z V(z) * H(z) + N \quad (2.6)$$

where  $H(z)$  is the on axis PSF calibrated at a depth  $z$  and  $V(z)$  is the set of scene points at a depth  $z$ . In other words,  $V(z)$  is the collection of irradiance values of all scene points exactly at a depth  $z$  and 0 otherwise. In our study, we discretize depth into discrete depth planes parallel to each other. For this case, 2.6 reduces to [7]:

$$y = \sum_k V(k) * H(k) + N \quad (2.7)$$

where  $k$  is the  $k$ th depth plane.

## 2.5 Summary of Assumptions

- 2D Forward model: The depth is taken to be greater than  $z_{max}$  or the scene is assumed to be 2D to ignore the depth scaling of PSF. The scene is assumed to be static such that the rolling shutter is processed as global shutter.

- Rolling shutter Forward model: The depth is taken to be greater than  $z_{max}$  or the scene is assumed to be 2D to ignore the depth scaling of PSF. The scene is assumed to be dynamic.
- 3D Forward model: The depth scaling of PSF is taken into consideration. The scene is assumed to be static such that the rolling shutter is processed as global shutter.
- The sensor is large enough to ignore cropping effects.
- The PSF is shift invariant.



## CHAPTER 3

### Related work

#### 3.1 Mask Based Lensless Imaging

Ultra-thin mask-based lensless cameras replace the lens of a traditional lens-based system with an optical mask typically placed close to the sensor. FlatCam[11] uses a separable amplitude mask placed approximately a millimetre from the sensor that was used to show 2D imaging and 3D volume reconstruction in [5]. DiffuserCam[6] used a random off-the-shelf diffuser as a mask placed 10 millimetres from the sensor. The authors demonstrated 3D imaging ability using this prototype. More recently, its ability to do high speed imaging[10] has also been demonstrated. PhlatCam[7] 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. As mentioned previously we will be using the same forward model and PSFs as [7].

#### 3.2 Learning for Lensless Imaging

Recently many learning based algorithms have been proposed for various types of lensless scene reconstructions. [1] proposed a feed forward deep network that performed photorealistic 2D scene reconstructions from separable mask FlatCam measurements utilizing a learnable fourier based inversion. [2] proposed an unrolled deep network for performing 2D image reconstructions from DiffuserCam measurements. Recently, [3] proposed FlatNet that was shown to perform 2D intensity reconstructions for any general lensless system, for both separable and non-seperable masks. [4] proposed a deep image prior based unsupervised method for lensless reconstructions.

### 3.3 High Speed Lensless Imaging

Traditionally, videos that are captured by sampling the entire grid of pixels for each frame are compressed by exploiting spatial and temporal redundancies. Compressive video sensing aims to exploit these redundancies during the capture itself. This can be used to overcome the chip bandwidth limit that constrains conventional cameras. the Rolling shutter Lensless system is essentially a low cost system for compressive sensing where the PSF provides global spatial multiplexing and the CMOS rolling shutter function provides a natural temporal encoding that is essential for compressive sensing. [10] uses an iterative algorithm FISTA to recover 140 frames from a single DiffuserCam capture. However FISTA requires highly compressible and sparse scenes to work well. Even for 2D image recovery from PhlatCam and DiffuserCam measurements, FISTA fails to generalise on complex scenes and hence learning based algorithms were developed for reconstruction of 2D scenes(specified in the section above). The only learning based algorithm existing in literature for lensless reconstruction was proposed by [4]. This algorithm utilizes an unsupervised deep image prior to reconstruct frames from a DiffuserCam measurement. The algorithm requires about 60,000 iterations per image during inference time which requires several hours on a typical GPU. The results are also over smoothed and lack temporal consistency. Our work proposes a Feed Forward Network to reconstruct frames from a single measurement that aims with higher perceptual quality in a fraction of a second. Further we extend our framework for high speed motion recognition from a single rolling shutter measurement using a similar feed forward approach.

### 3.4 Motion Recognition

Object video recognition has many interesting applications in fields like surveillance, gaming and assisted living environments. Many approaches exist in literature for video classification tasks and motion recognition, however due to the limited bandwidth of traditional lensed cameras, highspeed motion cannot be discerned. As a solution to this, the rolling shutter measurements can yield motion information with the same disadvantages of shutter and focusing action of lenses. However The global multiplexing of

lensless images opens up a new possibility of accurate motion recognition from rolling-shutter lensless measurement. In this work, we propose an end-to-end framework for predicting human motion classes using simulated rolling shutter lensless measurements. Further we also compare it with a corresponding high speed lensed camera aka the "best case classifier" using a well known comprehensive dataset [12].

### 3.5 3D Lensless Imaging

There are no deep learning 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.[6, 7] performed 3D voxel reconstructions from 2D lensless measurement using strong scene priors with iterative optimization routines like FISTA and ADMM. The authors in [8] proposed a joint intensity and depth reconstruction framework using an alternating optimization algorithm. However, this approach was only shown for a separable model and extending it to non-separable model is non-trivial since memory required quickly blows up. The above works rely on strong hand-crafted priors, traditional optimization routines and are iterative in nature. In contrast, our proposed approach is based on a feed-forward neural network that learns the priors from the data itself and is extremely fast. [13, 14] use programmable masks and multiple measurements for 3D estimation from thicker form-factor lensless cameras. We show 3D imaging capabilities of lensless cameras for a single passive mask with an ultra-thin geometry which is a much more challenging problem.





## CHAPTER 4

### Proposed Method

#### 4.1 Highspeed video reconstruction from single rolling shutter lensless captures

Our framework comprises of three parts, the MeasNet, the fourier based trainable inversion layer and a CNN based reconstruction block called the EnhanceNet.

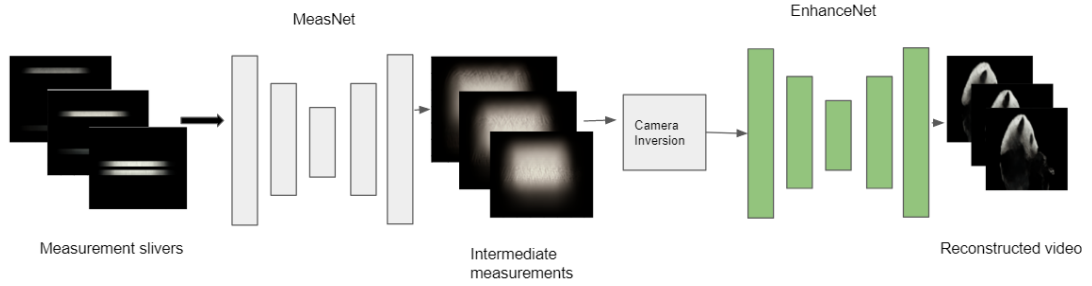


Figure 4.1: **Highspeed Reconstruction framework** Our proposed network first maps the measurement slivers into an intermediate measurement stack. The measurement stack is then deconvolved and passed through a convnet for perceptual enhancement of video.

##### 4.1.1 MeasNet

Since we only have access to a single rolling shutter measurement, we need to recover frame-specific information for a given timestep and shared information regarding the entire video. We achieve this by multiplying the rolling shutter measurement with the shutter function or equivalently selecting the exposing rows for a particular time step. We refer to these as the measurement-slivers. The aim is to reconstruct the lensless measurement corresponding to each frame which we refer to as the intermediate measurements. The intermediate measurements form a video of lensless images wherein each image is the 2D global shutter measurement corresponding to each scene frame.

Due to its capabilities in image reconstruction, we employ a Unet[15, 16] to achieve this extrapolation. MeasNet allows us to bypass the need for a RNN which requires huge amounts of memory and computation power, especially for high frame rates.

#### 4.1.2 Trainable Inversion

We base our Trainable inversion layer on [3]. The inversion layer performs the following operation individually for each frame:

$$\hat{S} = \mathcal{F}^{-1}(\mathcal{F}(W) \odot \mathcal{F}(Y)). \quad (4.1)$$

where  $Y$  is the intermediate measurement predicted by MeasNet,  $\hat{S}$  is the deconvolved Image,  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  are the DFT and the inverse DFT respectively,  $\odot$  is the Hadamard product or the element-wise product and  $W$  is a trainable filter initialized by the well-known Wiener filter:  $\mathcal{F}^{-1}(\frac{H^*}{\gamma + |H|^2})$  where  $\gamma$  is a regularization constant.

#### 4.1.3 EnhanceNet

Since the intermediate outputs can have errors and since the Fourier deconvolution is extremely sensitive to such differences, the deconvolved outputs are passed through a perceptual enhancement stage based on a Unet[15, 16]. This Unet is used to map the intermediate deconvolved outputs to perceptually enhanced video frames.

#### 4.1.4 Loss Function

**MSE Loss:** The Mean Squared error is used to measure distortion between the ground truth and the estimated intensity image. Given the ground truth image  $I_{true}$  and the estimated image  $I_{est}$ , this is given as:

$$\mathcal{L}_{MSE} = \|M_{gt} - I_{pred}\|_2^2 = \sum_{n=0}^N (I_{true,n} - I_{est,n})^2. \quad (4.2)$$

If  $M_{gt}$  and  $M_{pred}$  are the ground truth and predicted intermediate measurements then:

$$\mathcal{L}_{MSE-meas} = \|M_{gt} - M_{pred}\|_2^2 \quad (4.3)$$

If  $I_{gt}$  and  $I_{pred}$  are the ground truth and predicted frames then:

$$\mathcal{L}_{MSE-img} = \|I_{gt} - I_{pred}\|_2^2 \quad (4.4)$$

**Perceptual loss:** we use perceptual loss introduced by Johnson et al [17]. We use a pre-trained VGG-16 [18] model to extract feature-maps between the second convolution and second max pool layer and between the third convolution layer and fourth max pool layer. Suppose these activations are denoted as  $\psi_{22}$  and  $\psi_{34}$ , then

$$\mathcal{L}_{percept} = \|\psi_{22}(I_{gt}) - \psi_{22}(I_{pred})\|_2^2 + \|\psi_{34}(I_{gt}) - \psi_{34}(I_{pred})\|_2^2 \quad (4.5)$$

**Total loss:** The total loss is given by a weighted sum of the MSE and perceptual losses as

$$\mathcal{L}_{total} = \lambda_{percept}\mathcal{L}_{percept} + \lambda_{image}\mathcal{L}_{MSE-img} + \lambda_{meas}\mathcal{L}_{MSE-meas} \quad (4.6)$$

where  $\lambda_{percept}$ ,  $\lambda_{image}$  and  $\lambda_{meas}$  are the weighting constants.

## 4.2 Motion recognition from single rolling shutter lensless captures

We base the framework for motion recognition from rolling shutter lensless captures on our approach for image reconstruction. This network has a MeasNet and a Trainable inversion exactly as the ones above that are used to produce the intermediate RGB outputs.

### 4.2.1 ClassificationNet

Instead of passing the intermediate RGB outputs to the EnhanceNet for perceptual enhancement, we directly pass them to a CNN based feed-forward classification network. This also reduces the overall model complexity introduced by the bulky Unet of the EnhanceNet. This network can be a 2D or a 3D CNN or a special network like Resnet-18. The output of this network is a probability vector (after applying the softmax activation) corresponding to 10 classes used for the Crossentropy loss and argmax predictions.

### 4.2.2 Loss Function

**Crossentropy loss:** For the multiclass classification problem, the categorical crossentropy loss is given as

$$\mathcal{L}_{CCE} = - \sum_{c=1}^M y_{o,c} \log(p_{o,c}) \quad (4.7)$$

where  $y_{o,c}$  is the binary indicator that the ground truth values belong to a class  $c$  and  $p_{o,c}$  is the probability vector obtained from the softmax activation of the model output.  $M$  is the total number of classes.

**Total loss:** The total loss is given by a weighted sum of the MSE and CCE loss as

$$\mathcal{L}_{total} = \lambda_{image} \mathcal{L}_{CCE} + \lambda_{meas} \mathcal{L}_{MSE-meas} \quad (4.8)$$

where  $\lambda_{image}$  and  $\lambda_{meas}$  are the weighting constants.

## 4.3 3D Reconstruction from single lensless measurements

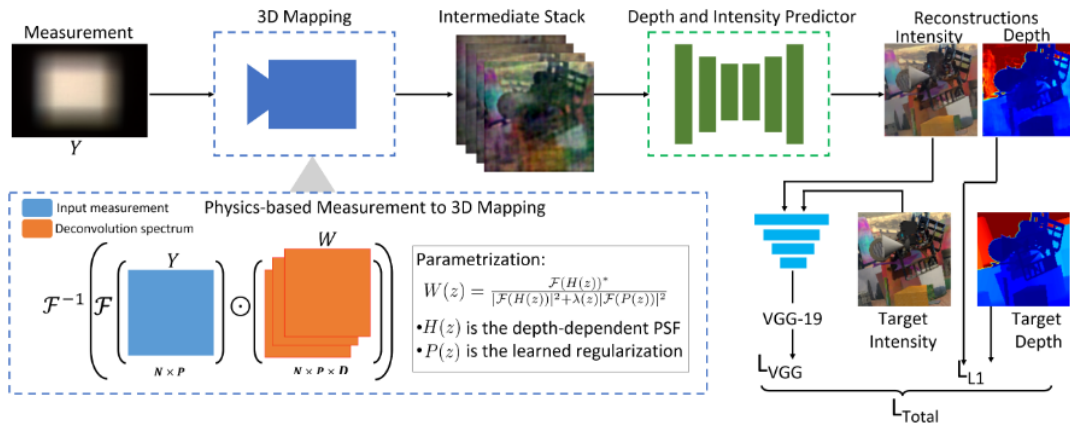


Figure 4.2: **FlatNet3D**. Our proposed network first maps the measurement into an intermediate 3D stack. A convnet then uses this stack to generate intensity and depth estimates. Finally, the entire network is trained in an end-to-end fashion using VGG loss on intensity images and L1 loss on depth maps.

### 4.3.1 Physics-based measurement to 3D mapping

Lensless measurements, due to global multiplexing, lack local structures[3, 2]. Also, from Equation 2.7 it can be seen that the measurement is a compressed representation of the 3D volume. Hence, we need to map the measurement to an intermediate stage having 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 formulation:

$$S_E(z) = \arg \min_{S(z)} \|Y - H(z) * S(z)\|_F^2 + \lambda(z) \|P(z) * S(z)\|_F^2. \quad (4.9)$$

where,  $*$  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. The solution to Equation 4.9 can be represented as,

$$S_E(z) = \mathcal{F}^{-1}(\mathcal{F}(Y) \odot W(z)), \quad (4.10)$$

where,

$$W(z) = \frac{\mathcal{F}(H(z))^*}{|\mathcal{F}(H(z))|^2 + \lambda(z)|\mathcal{F}(P(z))|^2} \quad (4.11)$$

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

Given that both Equations 4.10 and 4.11 are fully differentiable, we learn the filters  $P(z)$  and the vector  $\lambda(z)$ . Unlike the trainable inversion stage of FlatNet[3] as well as the highspeed framework, the learned mapping corresponding to FlatNet3D has much less parameters. This is because of the efficient parameterization of  $W(z)$  using Equation 4.11, where only  $P(z)$  and  $\lambda(z)$  are learned, which are of much lower dimensions.

### 4.3.2 3D stack to intensity and depth prediction

The next stage of FlatNet3D takes in  $S_E$  from the previous stage and outputs the depth and intensity predictions. Owing to its large scale success in image-to-image translation, image segmentation and depth reconstruction problems, we choose a U-Net [15]

to learn this mapping from  $S_E$  to final intensity and depth map. We have kept the kernel size fixed at  $3 \times 3$  and the number of filters have been gradually increased from 64 to 1024 across 5 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 4 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 [19] that enables the model to “focus” on the important regions and retain only the necessary activations.

### 4.3.3 Loss Function

**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  $d_{pred}$

$$\mathcal{L}_{L1} = ||d_{ref} - d_{pred}||_1 \quad (4.12)$$

**VGG loss:** To learn the intensity mapping, we use the VGG loss proposed in [17] which is known to produce photorealistic images. Let  $\phi_j$  denote the feature map of size  $C_j \times W_j \times H_j$  obtained by the j-th activation within the VGG19 network. The 22nd convolutional layer in 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  $I_{ref}$ .

$$\mathcal{L}_{VGG} = \frac{1}{C_j W_j H_j} ||\phi_j(I_{rec}) - \phi_j(I_{ref})||_2^2 \quad (4.13)$$

**Total loss:** The total loss used for joint training is given by a weighted sum of the L1 and the VGG loss as

$$\mathcal{L}_{total} = \mathcal{L}_{VGG} + \alpha \mathcal{L}_{L1} \quad (4.14)$$

where  $\alpha$  is the weighting constant.

# CHAPTER 5

## Experiments and Results

### 5.1 Implementation Details

For **Lensless Video Reconstruction**, We use the labeled part of the YouTube-VOS dataset [20]. Since the scenes are extremely complicated and have camera jitter, we remove the background using the ground truth video labels and filter out only the salient objects. By doing this, we are ensuring that the camera is stable and only the scene is varying while simulating. This dataset consists of 3420 train 49 validation samples. Since the number of frames in the dataset is less than what is required for simulating the lensless measurements, we use SuperSlomo [21] interpolation to increase the frame-rate by 6 times. We add Gaussian noise with a standard deviation of  $5 \times 10^{-3}$  to the final measurements. For our experiments, we chose a scene size of  $160 \times 160$  and a measurement size of  $320 \times 384$ . We set the number of measurement-slivers as 10, as a result of which our framework provides 10 intensity frames per lensless rolling shutter measurement. Following Antipa et al. [10], we use a dual rolling shutter. The model is trained using Adam [22] optimizer with initial learning rate of  $3 \times 10^{-4}$  with cosine annealing scheduler[23]. In our experiments, we set  $\lambda_{MSE_{meas}}$ ,  $\lambda_{MSE_{img}}$  and  $\lambda_{percept}$  to be 1, 5 and 0.1 respectively. We use both 2D [15] and 3D Unets [16] for both MeasNet and EnhanceNet where the number of filters is gradually increased from 32 to 128 and then decreased back to 32. NVIDIA GeForce GPUs were used for our experiments.

For **Lensless Activity Recognition**, we use the Something-Something dataset (Version V2) [12]. The Something-Something dataset consists of humans interacting in simple actions with everyday objects, and we choose the dataset for its diversity. The train set and validation set are constructed from 28,000 and 4900 randomly sampled videos from the dataset. For this dataset also, the number of frames in a video is lesser than the number required by the shutter and hence we interpolate the frames to yield videos having 4 times the frame-rate using SuperSlomo[21]. The number of classes in this dataset is 174 and to make the classification task simpler, we use filter the same



41 easy classes and map them to the same 10 coarse classes defined in [12]. Finally, we add Gaussian noise with a standard deviation of  $5 \times 10^{-3}$ . For our experiments, we chose a scene size of  $160 \times 160$  and a measurement size of  $320 \times 384$ . Following Antipa et al. [10], we use a dual rolling shutter. We set the number of measurement-slivers as 10, as a result of which our framework provides 10 intensity frames per lensless rolling shutter measurement. We use the 2D Dronet[24] as the ClassificationNet. In our experiments, we set  $\lambda_{MSE_{meas}}$  and  $\lambda_{CCE}$  to be 1, and 3.0 respectively. The model is trained using Adam [22] optimizer with initial learning rate of  $3 \times 10^{-5}$  with cosine annealing scheduler[23]. NVIDIA GeForce GPUs were used for our experiments.

For the **3D Lensless reconstruction task** we consider a scene depth range of upto 20cm. We capture real PSFs by placing a point source at 25 different depths between 3.6cm to 20cm from PhlatCam. Since there is no existing dataset for lensless depth and intensity with measurements and groundtruth pairs, we utilize existing RGB-D datasets for generating simulated measurements. Hence, we use intensity and disparity images of a subset of the FlyingThings3D[25] dataset to generate lensless measurements using the forward model described in Equation 2.7. We use 26066 RGB-D scenes for this purpose(21818 for training, 3000 for validation and 1248 for testing). Given that the FlyingThings3D dataset provides only the disparity, we first convert it to depth and scale the depth to the above range. Using the captured PSFs, the scenes, and the depth maps, we simulate the measurements. Finally, we add Gaussian noise to the simulated measurement so that the measurement PSNR corresponds to 20-50dB. We then use this simulated dataset for training and testing purpose. In our experiments, we consider a scene size of  $128 \times 128$ . For testing on real data, we used real PhlatCam captured with scenes placed within the above depth range in front of the camera. The Adam optimizer[26] 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  $\alpha$  was varied from 0.001 to 0.002. NVIDIA Titan X GPUs were used for our experiments.

## 5.2 Baselines

### 5.2.1 Highspeed Lensless reconstruction

We compare our proposed method with some of the traditional iterative algorithms. Video reconstruction is achieved by solving a TV-norm regularised least squares problem:

$$\arg \min_{v \geq 0} \frac{1}{2} \|Av - b\|_2^2 + \tau \|\nabla_{x,y,z} v\|_1 \quad (5.1)$$

where  $v$  is the scene,  $A$  is the linear operator denoting the camera forward model,  $b$  is the lensless measurement,  $\tau$  is a weighting constant and  $\nabla_{x,y,z}$  is the gradient operator. [10] uses FISTA to solve this formulation of the problem and can recover 140 frames from each measurement.

Monakhova et al [4] use an untrained deep network (UDN) utilize a Unet-like network as an image prior. The network is optimised so that the estimated video is consistent with the forward model. The network is untrained, that is it does not have a training stage and requires a number of iterations during its test time. This testing phase is akin to the traditional training of neural networks but it is specific to that image only and hence the model has to be 're-trained' or 'iterated' for every image. Hence, this framework falls under the class of Deep-Image-Prior methods for image reconstruction.

### 5.2.2 Highspeed Activity recognition

To compare our approach, we take the classifying CNN and directly train it on the natural images corresponding to the input measurement-slivers provided by the something-something dataset. This provides a ceiling performance for our model that is similar to the baselines proposed in the dataset's paper[12].

### 5.2.3 3D Lensless reconstruction

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

given by:

$$\hat{S}(z) = \arg \min_{S(z)} \|Y - H(z) * S(z)\|_F^2 + \lambda \|L * S(z)\|_F^2. \quad (5.2)$$

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  $3 \times 3$  and  $\lambda$  is a constant. The solution to Equation 5.2 is given by:

$$\hat{S}(z) = \mathcal{F}^{-1} \left( \frac{\mathcal{F}(H(z))^*}{|\mathcal{F}(H(z))|^2 + \lambda |\mathcal{F}(L)|^2} \odot \mathcal{F}(Y) \right). \quad (5.3)$$

This is the commonly used 2D Constrained Least Squares (CLS) filter used in image processing [27]. The second approach is to use the alternating direction method of multipliers (ADMM) proposed in [6, 7]. The ADMM based approach treats the 3D reconstruction problem as a regularized least-squares optimization problem:

$$\hat{S} = \arg \min_S \frac{1}{2} \|Y - \sum_z H(z) * S(z)\|_2^2 + \lambda \|\Psi(S)\|_1 + \lambda_1 \|S\|_1 \quad (5.4)$$

Here,  $\Psi$  is the gradient operator,  $S$  is the 3D volume and  $\hat{S}$  is its estimate. Rest of the notations have the same meaning as that in Equation 5.2.

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 [28] to minimise the following energy function formulated in [29]

$$E(x) = \sum_{i \in \mathcal{V}} E_i(x_i) + \lambda \sum_{(i,j) \in \mathcal{E}} E_{ij}(x_i, x_j) \quad (5.5)$$

Where  $E(x)$  is the energy of a depth labelling  $x$ ,  $x_i$  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.  $E_{ij}(x_i, x_j) = |x_i - x_j|$  where  $(i, j) \in \mathcal{E}$ , the set of edges connecting adjacent pixels.  $\lambda$  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 compare against a modified version of FlatNet[3] which was originally proposed for 2D scene estimation from lensless measurements. We modify FlatNet by replacing its inversion stage with the learned mapping. 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 refer to this model as FlatNet2D.

### 5.3 Metrics:

For quantitative evaluation, we use a combination of PSNR (in dB), SSIM and LPIPS[30] 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 video classification we report the top-k accuracy for  $k \in [1, 2, 5]$ . Top-k accuracy can be interpreted as the fraction of times, the top-k predicted probabilities indeed belong to those k classes and hence the top-1 accuracy is basically the conventional model accuracy. For depth predictions, we use Root Mean Square Error (RMSE) (in cm) for our evaluation. We report the average metrics evaluated on the simulated test set only since the real data doesn't have corresponding groundtruths.

## 5.4 Results

### 5.4.1 Comparison with baselines

#### 5.4.1.1 Highspeed video reconstruction

We compare both our proposed method using a 2D Unet and a 3D Unet for two scenes in figures and with the traditional methods. We see that our proposed methods produces images with much better perceptual quality and temporal consistency for complex motions than FISTA and UDN. UDN suffers from over-smoothing which is reflected in

its high PSNR values which FISTA has points which fail to move even though the original scene has them moving. UDN takes 4-6hrs per image, FISTA takes about 10mins and our method takes a fraction of a second per image. Further we notice that the 2D model works better than the 3D model, this is because that 2D model can be made bigger without too much computational overhead and the enhancenet is initialized with Flatnet[3].

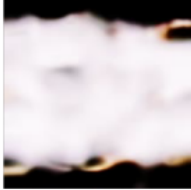
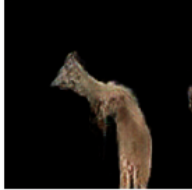

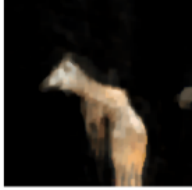
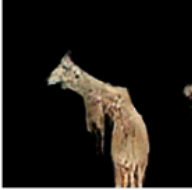
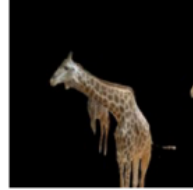
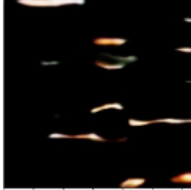
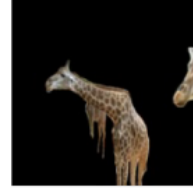
UDN 28.662/0.569	FISTA 31.92/0.267	Proposed-3D 19.88/0.227	Proposed-2D 20.19/0.162	Ground Truth PSNR/LPIPS
				
35.15/0.566	33.28/0.175	21.99/0.189	20.68/0.131	PSNR/LPIPS
				
33.98/0.587	32.09/0.221	20.89/0.185	21.62/0.117	PSNR/LPIPS
				
33.09/0.572	32.88/0.193	22.16/0.17	21.69/0.12	Mean PSNR/LPIPS

Figure 5.1: **Qualitative comparison on simulated Youtube VOS images.** Our method surpasses both the baselines in terms of perceptual quality and temporal consistency.

#### 5.4.1.2 Highspeed motion recognition

We compare our proposed method with the best case classification model using the same CNN classifier but training it using simply the natural images from the dataset without using the rolling-shutter measurements. The top-k accuracies are given in table 5.1 Additionally, we also compare them with the pre-trained 2D CNN metrics presented as a baseline in the original paper [12]. We can see that even in best case, the accuracies are quite low even for easy classes, this points to the extreme complexity of the problem that is reflected in poor accuracies.

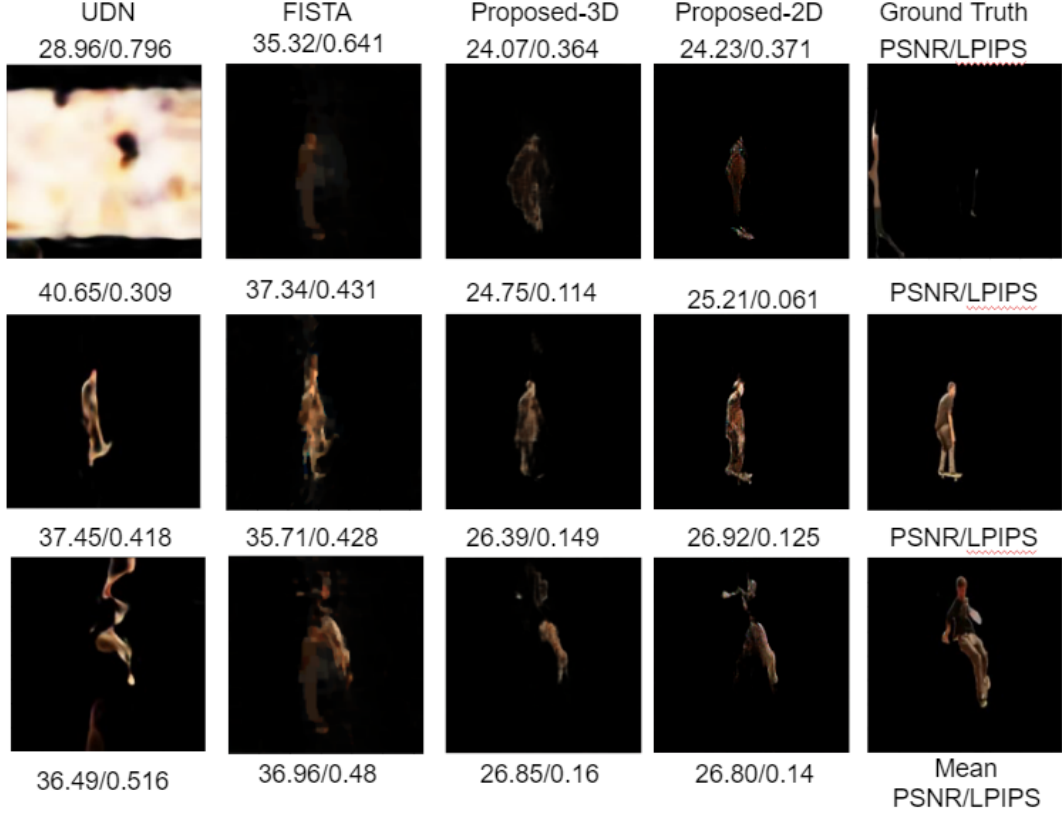


Figure 5.2: **Qualitative comparison on simulated Youtube VOS images.** Our method surpasses both the baselines in terms of perceptual quality and temporal consistency.

Method	Top-1	Top-2	Top-5
Proposed method	29.69%	46.92%	76.7%
DroNet[24]	39.13%	55.65%	82.55%
Pretrained-2D CNN [12]	45.30%	65.90%	N/A

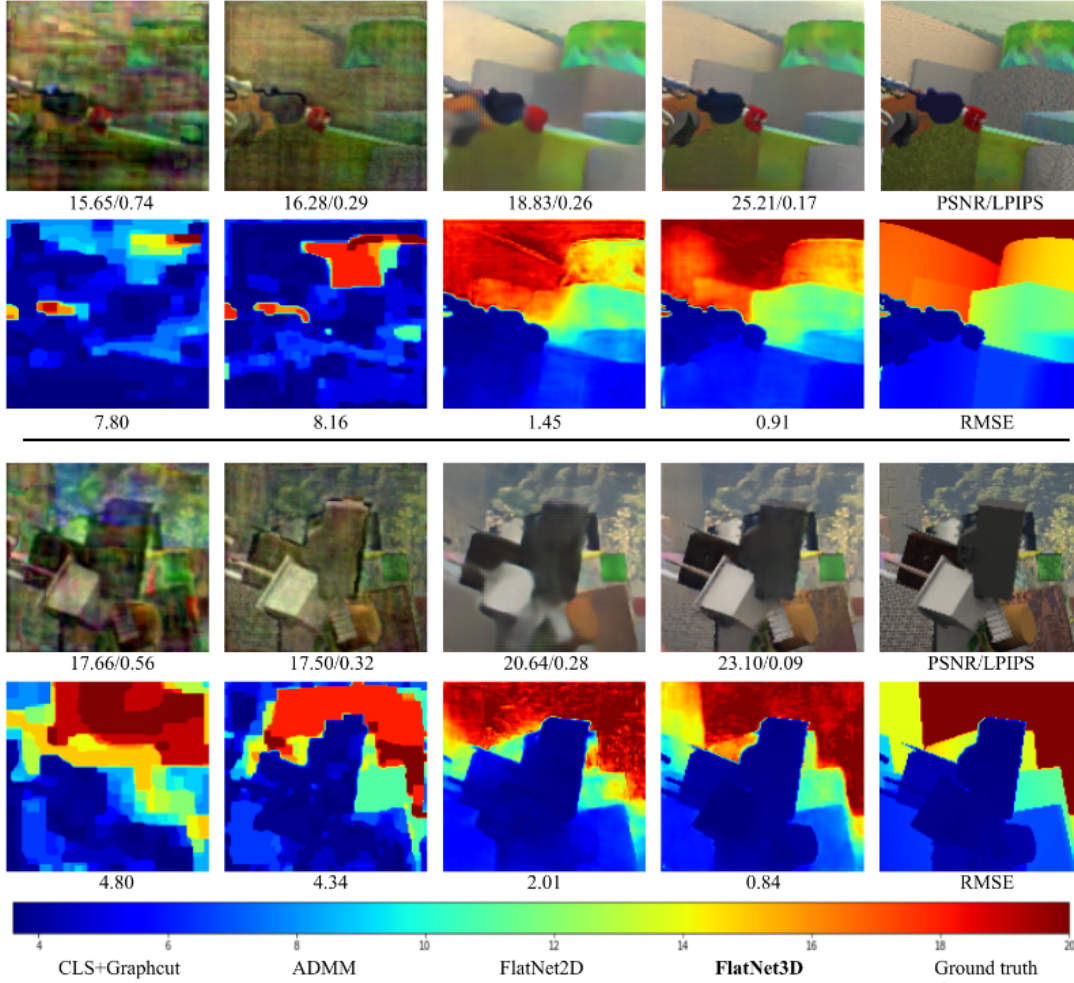
Table 5.1: **Comparison with the best case classifier** The accuracies here are calculated as  $100 \times (1 - \text{top-k error})$

#### 5.4.1.3 Simulated 3D lensless reconstructions

We present the quantitative comparison of our approach against the baselines. Table 5.2 reports the average metrics evaluated on the simulated test set described in section 5.1. We can see that FlatNet3D clearly outperforms traditional optimization based approaches like ADMM+Graphcut or CLS+Graphcut in terms of accuracy. Also, FlatNet3D is nearly 100x faster than ADMM+Graphcut and nearly 10x faster than CLS+Graphcut. Among the learning based approaches, FlatNet3D outperforms FlatNet2D because the latter isn't able to extract the depth information accurately from a single PSF. We provide visual results for the proposed FlatNet3D against the baselines in Figure 5.3 for two simulated scenes. We can conclude that FlatNet3D provides bet-

Method	PSNR (in dB)	SSIM	LPIPS	RMSE (in cm)	Inference Time (in sec)
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	<b>0.011</b>
<b>FlatNet3D</b>	<b>21.91</b>	<b>0.79</b>	<b>0.14</b>	<b>1.42</b>	0.013

Table 5.2: **Quantitative comparison with other approaches.** A comparison of the average metrics for the proposed FlatNet3D along with the baselines evaluated on the simulated test set.



ter depth maps and perceptual intensity quality which is reflected by lower LPIPS and RMSE values. The traditional approaches provide noisy intensity and depth estimates while FlatNet2D is unable to extract sharp all-in-focus image and accurate depth map from a single focal stack image.

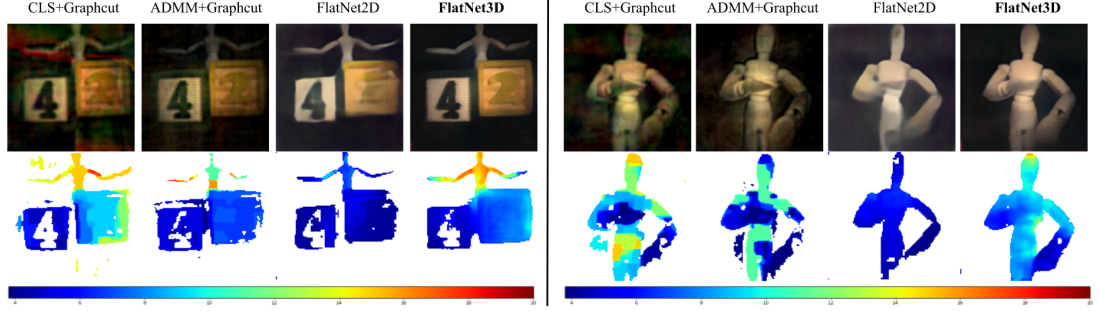


Figure 5.4: **Qualitative comparison on real captures.** We show real result for two scenes. FlatNet3D provides better contrast for intensity images and cleaner depth maps for both scenes.

#### 5.4.1.4 Real 3D lensless reconstructions

Since real data has no ground truth, we have provided a visual comparison on real data captured using PhlatCam[7] in Figure 5.4. Similar to results on simulated data, FlatNet3D provides better quality intensity and depth estimates. The depth estimates using FlatNet3D has fewer spurious regions especially for the right scene which lacks texture. Baseline traditional methods suffer from noise for these data as well, with ADMM+Graphcut performing better than CLS+Graphcut. As with the simulated case, 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 the figure 5.4 for better visualization.

### 5.4.2 Ablation experiments

In this subsection, we evaluate the performance of our 3D lensless model against its baselines for images having different noise levels. We vary the simulated measurement’s PSNR from 20-50dB 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 the validation dataset for each noise level separately for optimal performance. In Figure 5.5, we have shown the average LPIPS of intensity estimates and the RMSE of depth estimate for various noise levels. We can observe that traditional approaches perform consistently worse than the learning-based approaches both in terms of intensity and depth quality despite having carefully tuned parameters.



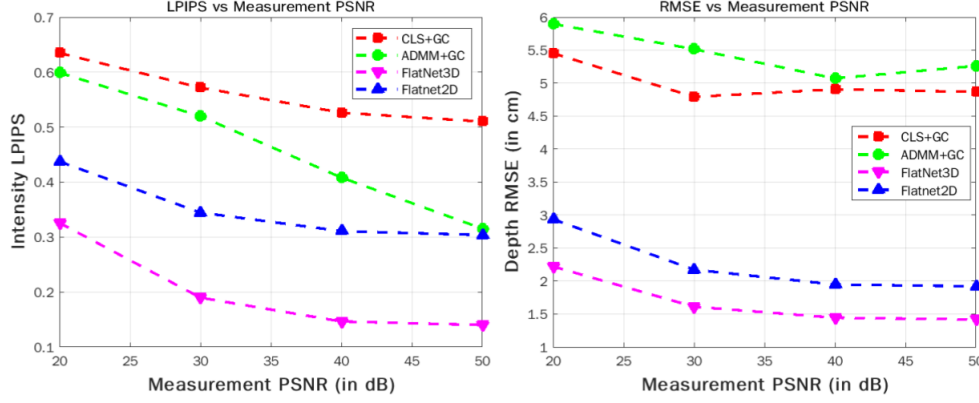


Figure 5.5: **Noise ablation.** We vary the measurement noise and evaluate the performance of all the methods.

### 5.4.3 Applications

Getting absolute depth information from endoscopic scenes is of vital importance for medical diagnosis of tumours, polyps and carcinomas. The primary difference between these images and the images of Flying Things 3D is that the scenes are extremely low in texture and that the surface is usually reflective due to mucus. 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 using our above 3D reconstruction paradigm. To do this, we simulate lensless measurements from the colon subset of the synthetic EndoSLAM dataset[31] 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 for testing. First, we have undistorted the input images using the calibration files provided by the authors. We have then scaled the depth to a maximum of 10cm and finetuned our trained FlatNet3D on this dataset. Figure 5.6 shows visual results for the intensity and depth maps for various scenes in the digestive tract. In spite of the having extremely low textures, FlatNet3D is able to extract absolute depth information from a single measurement.

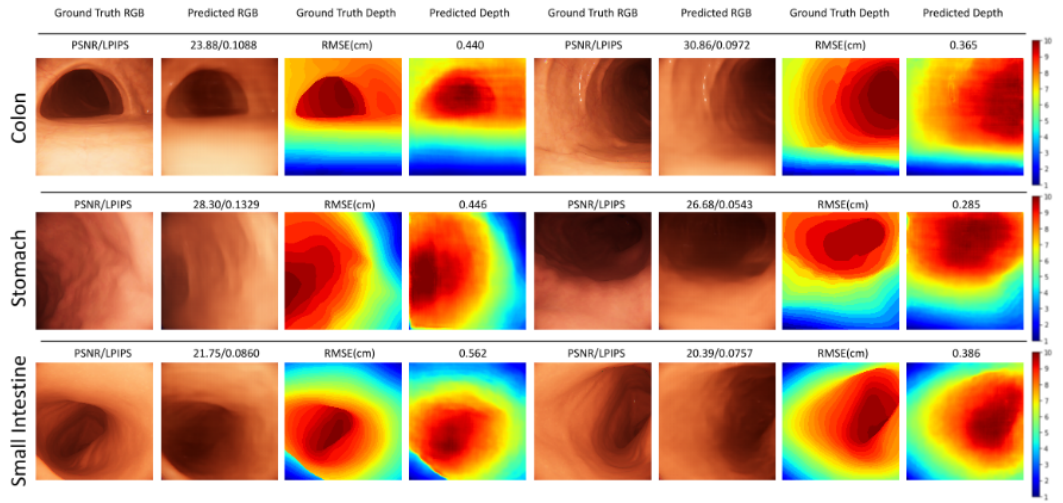


Figure 5.6: **Performance on EndoSLAM Dataset.** We have finetuned FlatNet3D on the EndoSLAM RGB-D dataset. It can be seen that FlatNet3D is able to provide high quality depth maps despite the scenes being extremely low in texture.



## CHAPTER 6

### Conclusion and Future work

In this work we propose an end-to-end novel deep network for reconstructing video from a single shot lensless capture. Our method produces sharper and perceptually better videos compared to the traditional methods. It is able to reconstruct videos in much lesser time(a fraction of a second) when compared to other methods(hours).

However our method need improvement on the temporal consistency that results from having lesser number of slivers and a relatively smaller CNN network. Also, the dataet used is complex but is still sparse. Further work can focus on better temporal consistency for non-sparse scenes. Recent literature shows promise in alleviating this, using optical flow based loss functions.

We also proposed an end-to-end feed forward network for motion recognition. The results shed light on the extreme complexity of the problem. Our approach cannot handle large motions although it shows comparable performance with the baselines. Further work can can focus on improving the model using recurrent networks and applying it to sparse point-based motion capture rather than dense activity recognition task.

We also 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 and the prior in the data for doing so, leading to superior quality intensity and depth estimates. A key component of our approach is the physics-based learned 3D stack mapping stage, which is very efficiently parameterized leading to less memory usage and computation.

FlatNet3D's performance is however, limited by the lensless camera's physical geometry constraints ( $z_{max}$ ). Although they can be leveraged for tasks within this depth range like monocular 3D endoscopy imaging and microscopy, it is unfit for 3D reconstructions beyond this depth range. Future work can explore the possibility of getting depth estimates using multiple lensless cameras, monocular cues and dynamic mask patterns similar to [13].



## REFERENCES

- [1] Salman S Khan, VR Adarsh, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7860–7869, 2019.
- [2] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Optics express*, 27(20):28075–28090, 2019.
- [3] Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [4] Kristina Monakhova, Vi Tran, Grace Kuo, and Laura Waller. Untrained networks for compressive lensless photography. *Optics Express*, 29(13):20913–20929, 2021.
- [5] Jesse K Adams, Vivek Boominathan, Benjamin W Avants, Daniel G Vercosa, Fan Ye, Richard G Baraniuk, Jacob T Robinson, and Ashok Veeraraghavan. Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science advances*, 3(12):e1701548, 2017.
- [6] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018.
- [7] Vivek Boominathan, Jesse Adams, Jacob Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [8] Yucheng Zheng and M Salman Asif. Joint image and depth estimation with mask-

- based lensless cameras. *IEEE Transactions on Computational Imaging*, 6:1167–1178, 2020.
- [9] E. O. Brigham and R. E. Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967.
- [10] Nick Antipa, Patrick Oare, Emrah Bostan, Ren Ng, and Laura Waller. Video from stills: Lensless imaging with rolling shutter. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019.
- [11] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.
- [12] R. Goyal, S. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thureau, I. Bax, and R. Memisevic. The “something something” video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society.
- [13] Yi Hua, Shigeki Nakamura, M Salman Asif, and Aswin C Sankaranarayanan. Sweepcam—depth-aware lensless imaging using programmable masks. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1606–1617, 2020.
- [14] Yucheng Zheng, Yi Hua, Aswin C Sankaranarayanan, and M Salman Asif. A simple framework for 3d lensless imaging with programmable masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2603–2612, 2021.
- [15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

- [16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.
- [17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.
- [19] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.
- [20] N. Xu, L. Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *ArXiv*, abs/1809.03327, 2018.
- [21] Huaizu Jiang, Deqing Sun, V. Jampani, Ming-Hsuan Yang, Erik G. Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.
- [22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
- [23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with restarts. *ArXiv*, abs/1608.03983, 2016.
- [24] Christos Kyrkou, George Plastiras, Theodoris Theodoridis, Stylianos I. Venieris, and Christos-Savvas Bouganis. Dronet: Efficient convolutional neural network detector for real-time uav applications. In *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 967–972, 2018.
- [25] 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.
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
  - [27] Bobby R Hunt. The application of constrained least squares estimation to image restoration by digital computer. *IEEE Transactions on Computers*, 100(9):805–812, 1973.
  - [28] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.
  - [29] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M. Seitz. Depth from focus with your mobile phone. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2015.
  - [30] 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.
  - [31] 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.