Learning Light Field Reconstruction from a Single Coded Image

Anil Kumar Vadathya[1], Saikiran Cholleti[1], Gautham Ramajayam[2], Vijayalakshmi Kanchana[3],
Kaushik Mitra[1]

[1]Computational Imaging Lab
IIT Madras, Chennai, India

[2]NIT Trichy
Tiruchinapally, India

[3]SSN College of Engg.
Chennai, India

{ee15s055,ee13b021,kmitra}@ee.iitm.ac.in

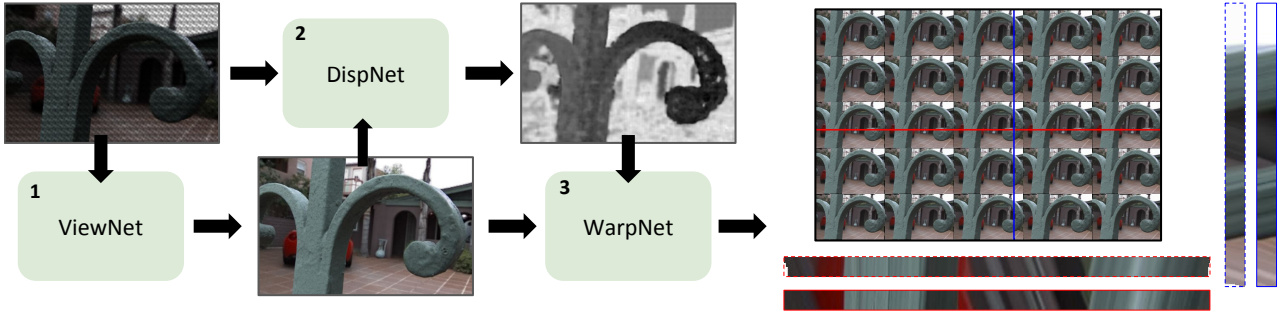gautham.1131@gmail.com

ajithabhavana@gmail.com

Figure 1: We propose a learning based solution for full sensor resolution light field reconstruction from a single coded image. Our approach consists of three convolutional neural networks (CNNs); *ViewNet* reconstructs center view from the coded image, *DispNet* estimates disparity map, *WarpNet* generates light field by warping center view with the disparity. As can be seen from EPI images, our approach recovers parallax very well. Generated EPI in dashed line border and GT in thick line.

Abstract

Light field imaging is a rich way of representing the 3D world around us. However, due to limited sensor resolution capturing light field data inherently poses spatio-angular resolution trade-off. In this paper, we propose a deep learning based solution to tackle the resolution trade-off. Specifically, we reconstruct full sensor resolution light field from a single coded image. We propose to do this in three stages 1) reconstruction of center view from the coded image 2) estimating disparity map from the coded image and center view 3) warping center view using the disparity to generate light field. We propose three neural networks for these stages. Our disparity estimation network is trained in an unsupervised manner alleviating the need for ground truth disparity. Our results demonstrate better recovery of parallax from the coded image and sharper reconstruction than dictionary learning approaches. All our results and code would be available at our project page [2].

1. Introduction

Imaging technologies have been striving to capture the rich three-dimensional scene around us as it is. But, since images conventional cameras have only been able to project the scene to a two-dimensional photograph. Light field imaging is a step closer in this direction. Light field captures different perspective shifts of the same scene. It enables post capture facilities like refocusing and view-point changes. This is realized by jointly modifying the camera optics and computationally processing the recorded data. The advent of light field imaging technology into affordable commercial cameras like Lytro [1] has renewed research interests resulting its applications in areas like cinematography, 3D imaging, AR and VR etc.

Many methods have been proposed for capturing light field data [16, 22, 3, 21, 8]. Of these, micro-lens array based acquisition by Ng et al. [21] has successfully been adopted into lytro camera [1]. Most of these methods suffer from a common problem. Due to limited sensor resolution trade-off arises between the spatial and angular resolution. These cameras have to sacrifice the spatial resolution to gain angular information. This hurts the megapixel-hungry trend with the consumer cameras.

Recently, high resolution light field imaging has received a lot of attention from the research community. Early attempts used computer vision techniques to enhance spatio-

1

angular resolution by exploiting the redundancy in 4D light field [8]. Compressive light field imaging works backed by compressive sensing (CS) theory attempt to reconstruct full sensor resolution light field either from a set of a coded images [4] or a single image [19]. Since the reconstruction is ill-posed, data priors are essential. With the success of deep learning techniques for low-level image processing, recently, some works have proposed to use it for high resolution light field imaging [13, 23].

In this work, we propose to use a deep neural network for reconstructing high resolution light field $\nu \times \nu \times h \times w$ from a single coded image $h \times w$. Our pipeline involves 1) center view reconstruction from the coded image 2) disparity map estimation from the center view and coded image and 3) warping center view using the disparity map to reconstruct light field. We show our results using simulations on Lytro illum dataset by Kalantari et al. [13]. Our major contributions are as follows:

- We propose a deep learning based method for full sensor resolution light field reconstruction from a single coded image.

- Output of our second network is disparity which in general would require ground truth disparity for training. Following the recent unsupervised learning works [9] we learn it in an unsupervised manner.

- Our disparity map based view synthesis inherently respects the parallax, depth dependent disparity in light field image. In contrast, a direct regression to light field from a coded image is not guaranteed to do so.

2. Related Work

Light Field Imaging: Light field imaging is a century old, initial attempts used a pinhole array in front of sensor to capture it, [7]. Later people used camera arrays and other hardware modifications [16, 8]. Being bulky and expensive they were not affordable for commercial use. Recently, lenset based light field imaging [21] has been integrated into commercial light cameras; Lytro [1] and Pelican Imaging [10]. This has renewed their application in 3D imaging. However, these lenslet cameras suffer from a common problem of spatio-angular tradeoff. They have to sparsely sample in either domain. For example, Pelican cameras have a 2×2 array.

High-resolution Light Field Imaging: Given the tradeoff between spatial-angular resolution many methods have been proposed to address the problem. Broadly these can be categorized in two categories. First set of methods sample the light field sparsely in one of the domains and post process using computer vision techniques to enhance spatio-angular resolution. Mitra et al. [20] proposed GMM light

field prior for light field super-resolution, Bishop et al. [5] propose a variational bayes framework for the same.

Second set of methods propose computational imaging based solutions. Levin et al. [15] sythesize 4D light field from 3D focal stack. Backed by compressive sensing, Marwah et al. [19] recover full sensor resolution light field from a single coded image. Babacan et al. [4] propose to reconstruct light field to reconstruct from a set of compressive measurements.

Compressive Light Field Imaging: Marwah et al. place a coded mask at a distance from sensor multiplexing the angular information to a coded image on the sensor. They use light field dictionary as a prior for reconstruction from the coded image. Instead, here we propose to use deep neural nets for reconstruction which are shown to outperform dictionary learning in many low-level image processing tasks (see figure 6).

Deep Learning: With the successful application of deep learning for low-level image processing. Learning based solutions have been proposed for first category of methods. Yoon et al. [23] do light field super-resolution by a factor of two using CNNs. Recently, Kalantari et al. [13] used learning based methods for view synthesis. Particularly, they were able to show convincing results on challenging real scenes by generating full 8×8 light field from only four corner views. In this work, we propose a deep learning solution for second category of methods. A learning based approach to reconstruct light field from coded image.

Unsupervised Disparity Estimation: Recently, unsupervised methods have been proposed for learning depth maps from stereo image pairs without access to ground-truth depth maps [6, 9]. Godard et al. [9] use photometric losses between the warped left view and right view achieving state of the art results on KITTI dataset. Unlike these methods, we don't have access to the stereo pair; only the coded image is used to estimate the depth.

3. Proposed Method

3.1. Coded Image Simulation

To simulate coded image we follow the same approach followed in [19]. Basically, this assumes a mask at a distance from sensor which optically modulates the incoming light field $L(x, \nu)$, projecting the coded image $I_c(x)$, on to the sensor. Mathematically this can be described as,

$$I_c(x) = \int_\nu f(x, \nu) L(x, \nu) d\nu, \qquad (1)$$

where, x is the spatial dimension on the sensor plane and ν is the angular dimension on the aperture plane. Given this coded image I_c, our goal is to recover the whole light field L, which is of the same resolution as the coded image. Thus, tackling the resolution trade-off.
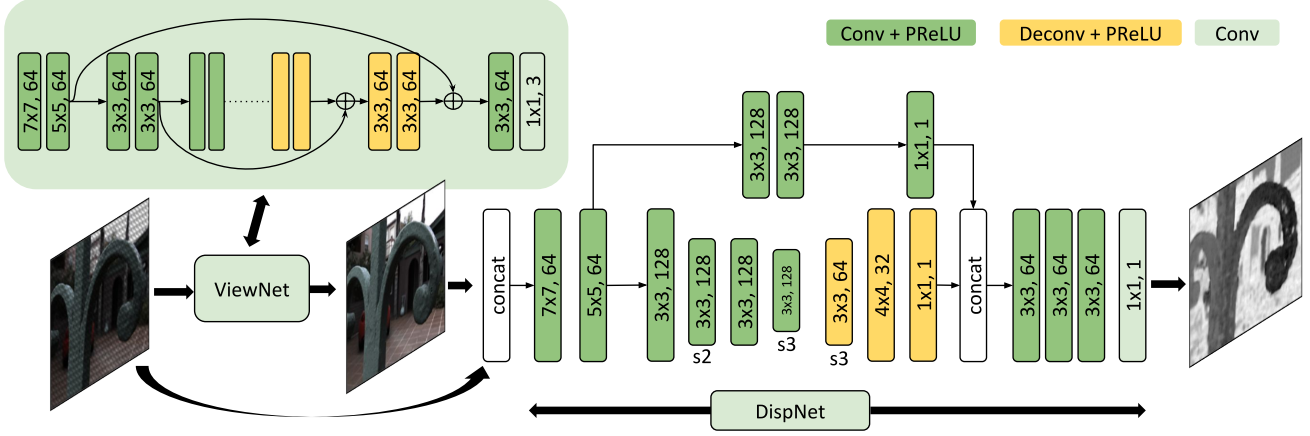
Figure 2: *ViewNet* and *DispNet*; figure shows the convolutional architecture details of both the networks. ViewNet is a convolutional deconvolutional architecture with three symmteric skip connections. It is shown compactly and can be easily inferred from Mao et al. [18]. DispNet is an encoder decoder (bottom) with a skip connection (top). The 'sk' specified below the block implies stride k convolution. Each block shows the filter size (c) and number of output channels (h) as $c \times c, h$.

3.2. DNNs for Light Field Reconstruction

A trivial way of doing this is to train a CNN to directly regress for novel views. This might be inefficient for two reasons. First, light field exhibits depth dependent disparity (epi-polar geometry) and direct regression to novel views is not guaranteed to respect this. Secondly, as is noted by recent works on view synthesis [6, 13] training a neural network directly to warp/translate image is inefficient. Hence, we break down the light field reconstruction into two steps of scene reconstruction (center view) and scene geometry (disparity) estimation, then render full light field. Figure 1 shows the proposed pipeline for compressive light field reconstruction.

In the following sections, we discuss the individual network architecture details.

3.2.1 Center View Reconstruction: ViewNet

We need to reconstruct one of the views to synthesize the novel views. Intuitively, center view (L_0) of the light field is ideal for this. We use a series of convolutional layers for reconstructing center view, L_0 from the coded image I_c. The regions further away from the focal plane are proportionately degraded in I_c as they see much parallax. To better reconstruct such details we incorporate symmetric skip connections proposed by Mao et al. [18] (see figure 2).

3.2.2 Learning Disparity Estimation: DispNet

Once we have the center view reconstructed from ViewNet, we use both the coded image, I_c and center view, L_0 to get the disparity map, D corresponding to the center view. Now, this disparity map can be used to generate novel view

at q, $\hat{L}_q$ from the center view L_0, under the Lambertian Scene assumption. This is mathematically given as,

$$L_q(x + qD(x)) = L_0(x), \quad (2)$$

where x is the pixel location. This is same as the forward warping in case of stereo pair.

Here, we propose a deep neural network, *DispNet*, for disparity map estimation. We stack the coded image and reconstructed center view as input to DispNet. Center view is used to ensure that disparity maps aligns with it.

DispNet shown in figure 2 is used for disparity estimation. It consists of encoder-decoder architecture with a skip-connection. This network is designed along the lines of recent unsupervised disparity estimation networks [9]. Encoder performs strided(> 1) convolutions encoding the disparity information. Also, this provides a way for increasing the receptive fields. Decoder performs strided deconvolutions bringing back the feature map to the input resolution. Skip connections ensure that the low level details are preserved in the disparity maps. The feature maps from decoder and skip connection are concatenated, followed by four convolutional layers to output final disparity map, D. This goes as input to the warping block.

3.2.3 Warping and Interpolation: WarpNet

The disparity map, D is used to synthesize the novel views by forward-warping shown in eq. (2). However, the presence of depth edges always introduces some occluded regions which can't be recovered by forward-warping. To tackle this, we use four more convolutional layers on top of the warped view for interpolation job. We refer to this network as *WarpNet* and it outputs the final novel view.

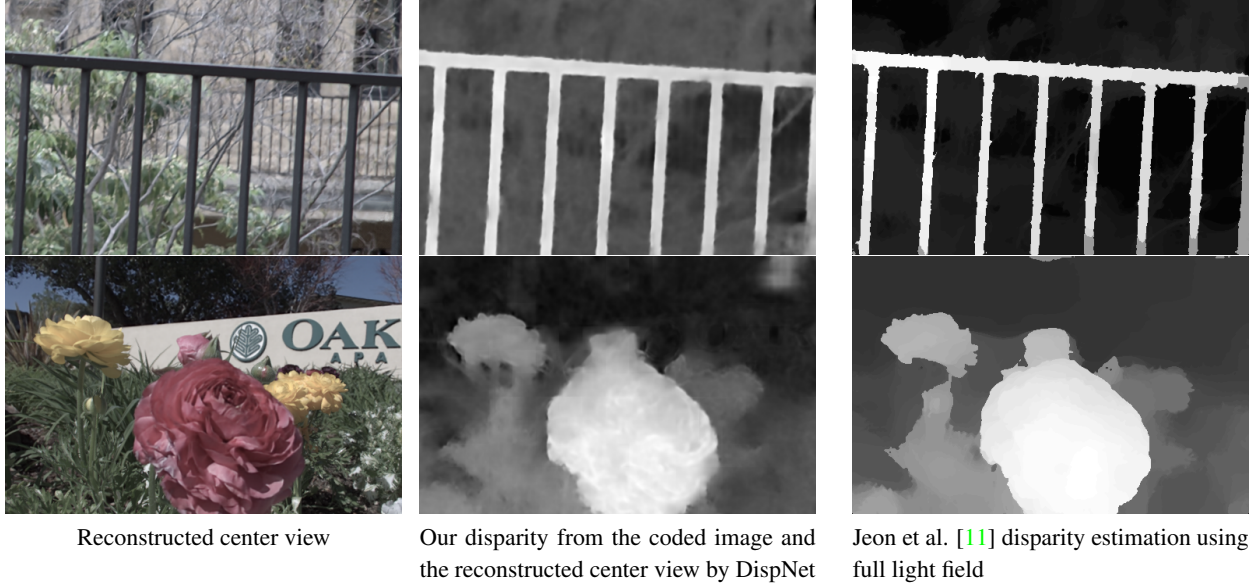| Reconstructed center view | Our disparity from the coded image and the reconstructed center view by DispNet | Jeon et al. [11] disparity estimation using full light field |

Figure 3: Disparity map comparison: Using only photo-consistency loss our network is able to learn disparity maps as good as one estimated from ground truth light field. Note that our estimation is without any explicit regularization unlike [11]. Please see our project page [2] for comparisons on a variety of scenes.
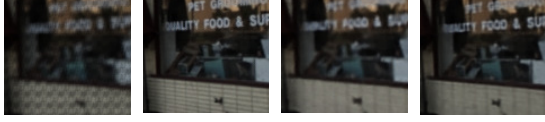

Figure 4: Center view reconstruction with *ViewNet*: Notice recovery of sharp details like text over the glass and pattern edges below it with Conv-Deconv network. Conv-Deconv network employs symmetric skip connections proposed by Mao et al.

4. Training

Finally, the generated view is compared with ground truth and loss is back-propagated through the warping and disparity estimator network. The loss is given as,

$$E = \sum_k (\hat{L}_{q,k} - L_{q,k})^2, \qquad (3)$$

where sum is over RGB color channels. The generated view at q, $\hat{L}_q$ is a function of parameters of both disparity estimator (W_d) and warping network (W_w).

$$\frac{\partial E}{\partial W_d} = \sum_k \left(\frac{\partial E}{\partial \hat{L}_{q,k}} \frac{\partial \hat{L}_{q,k}}{\partial D} \right) \frac{\partial D}{\partial W_d} \qquad (4)$$

For training using gradient descent we need the gradients of the loss (3) with respect to the parameters W_d, W_w. The gradients $\partial E / \partial W_d$ involves estimating the second term of

equation (4) which is gradient of warping function with dispartiy map, D. Although bicubic interpolation is differentiable numerical estimation of gradients gives satisfactory results. Other parameter gradients can be obtained using standard back-propagation techniques.

4.1. Implementation Details

For training, we use the 100 Lytro illum images provided by Kalantari et al. [13]. We split this data randomly into 85 training images and 15 validation images. We extract 2 million light field patches ($7 \times 7 \times 120 \times 120$) from the training data. We used random Gaussian code 15×15, clipped between 0 and 1 to simulate coded image (120×120). We use caffe [12] to train with Adam optimizer [14]. We set initial learning rate $1e-4$ and reduce it by a factor of 0.8 every 5 epochs. Except with the intial 7×7, 5×5 convolutional layers we do padded convolutions to preserve the input-ouput size. To accommodate for the loss in spatial resolution due to initial unpadded convolutions we appropriately crop the label (110×110) to evaluate the loss.

Initially, we pretrain the individual blocks separately and then combine them for an end-end training. ViewNet is individually trained for center view reconstruction. DispNet and WarpNet are trained together with ground truth center view initially. Pretraining is done for 30 epochs. In end-end training, the ViewNet output is used as input to the DispNet and warping function. This joint optimization is carried on for 5 epochs.

Ours   GT

27.94 dB, 0.935

26.12 dB, 0.887

Seahorse scene - 33.16 dB, 0.953

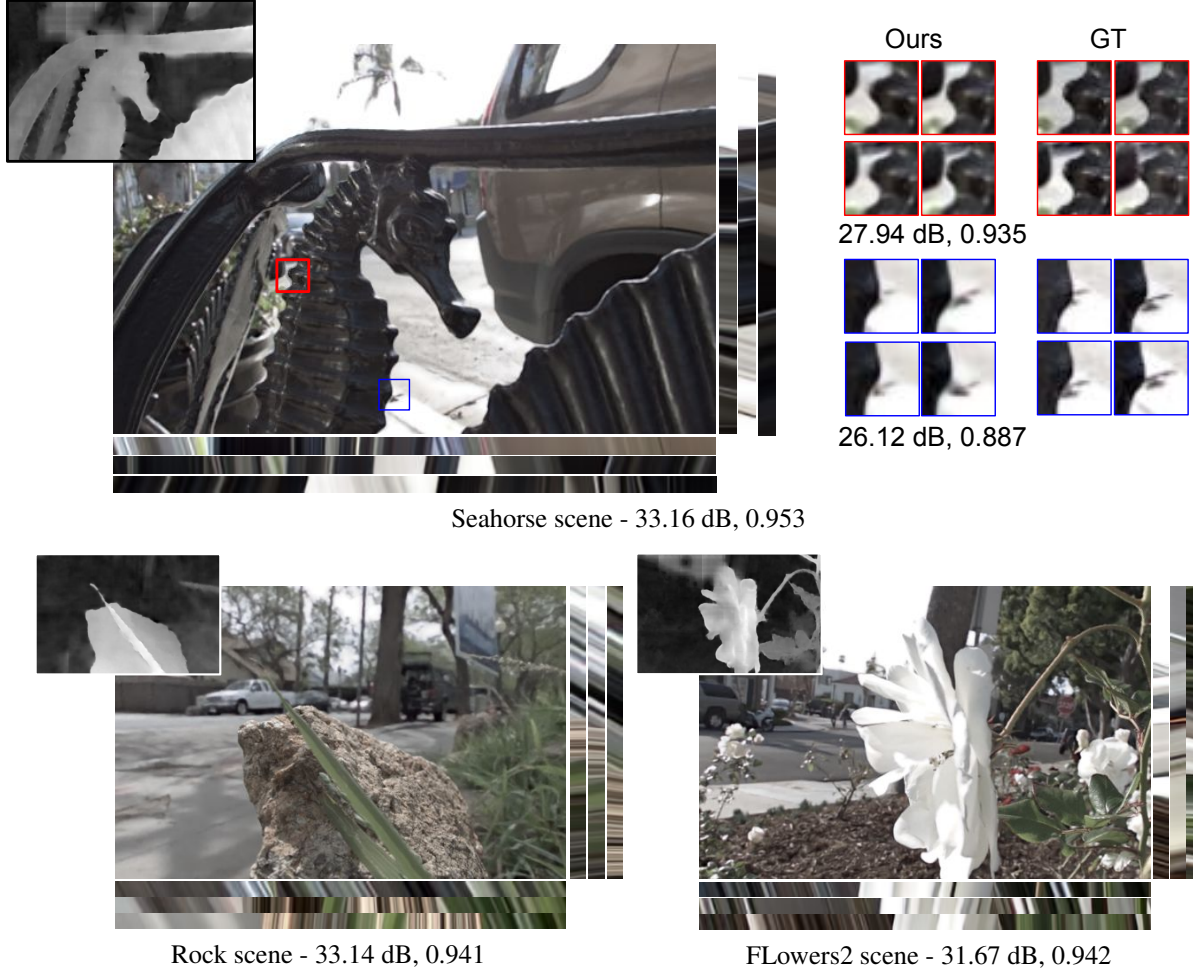Rock scene - 33.14 dB, 0.941   FLowers2 scene - 31.67 dB, 0.942

Figure 5: Full sensor resolution light field reconstruction from coded image for different test scenes from Kalantari et al. [13]: figures show the reconstructed center view overlaid with the estimated disparity maps (brigter is closest). The numbers reported are PSNR in dB and SSIM. From the slope of EPI images we can see that our approach recovers parallax well also implied from the sharpness of the disparity maps at occlusion edges. This can also be noticed from the insets (blue and red) showing the corner views of 7x7 light field of Seahorse scene. We can see that ViewNet recovers fine details like the shadow pattern in the blue inset. Please check our project page [2] for animations of generated lightfield clearly showing the parallax.

5. Results

Center View Reconstruction - As shown in figure 4, adding skip connections in viewnet helps in better recovery of finer details over a series of convolutional layers. Notice the details like the text on glass and objects behind the glass.

Disparity Estimation - Figure 3 shows DispNet estimation when trained with coded image and ground truth center view. We compare our disparity map with one estimated from full light field by Jeon et al. [11]. We can see that our method estimates disparities as good as them with out any explicit regularization. Figure 5 shows the disparity maps estimated from the coded image using DispNet. Note that we don't use any ground truth disparity for training Disp-

Net. We use the photoconsistency loss to learn the disparity map as has been recently employed for unsupervised disparity estimation [9]. From figure 5, we can see fine details in the disparity and its sharp around the occlusion edges, thus enabling better parallax recovery in light field.

Compressive Light Field Reconstruction - Figure 5 shows the results of light field reconstruction at 7x7 angular resolution from the coded image using our pipeline. It shows the reconstructed center view along with estimated disparity overlaid on it. We can see that reconstructed EPI are sharp. The mentioned PSNR and SSIM values are the average across all the 49 reconstructed views. Note the parallax recovery also shown in insets for the Seahorse scene.

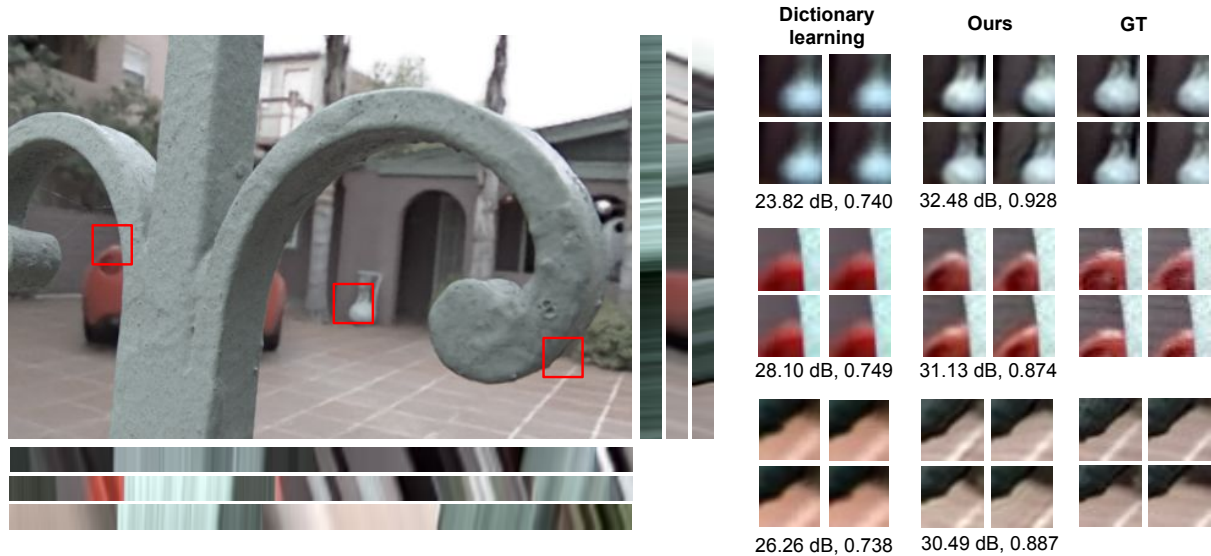**Comparison with Dictionary Learning** - We compare

Figure 6: Comparison with dictionary learning: figure shows the center view reconstructed from our approach and EPIs from the light field (5x5). For the patches indicated in the figure, we compare our reconstructions with that dictionary learning based method of Marwah et al. [19]. As can be seen, our method recovers the parallax well and sharper reconstruction.

Figure 7: Refocusing application: The Flowers scene is refocused using the estimated disparity and generated light filed. The focus shifts from front to back from left to right. Note that refocusing is done by shifting and adding the light field images not by disparity based blurring.

our learning based light field reconstruction to that of dictionary learning based method of Marwah et al. [19]. To reduce computational burden here we compare our light field synthesis at 5×5 angular resolution instead of 7×7. We retrain our network for this case. For dictionary learning, we extract $5 \times 5 \times 11 \times 11$ light field patches from 20 illum training images. After coreset reduction we obtain 40,000 patches for training. We use SPAMS toolbox for learning dictionary [17] and perform reconstruction along the lines of Marwah et al. Figure 6 shows comparisons with our method. We can clearly see that our approach recovers parallax much better than dictionary learning. Also, recon-

structions are sharper especially when the objects are farther from the focal plane like the pot in the background (top patch), the grid lines on the floor (bottom patch).

Refocusing - As an application of the reconstructed light field we show refocusing of the *Flowers* scene using the generated disparity map (see figure 7).

6. Conclusions and Future work

In this work we propose a learning based solution for sensor resolution preserving light field reconstruction from a single coded image. Our approach with explicit disparity

based reconstruction is suitable for parallax recovery in the light field. We learn disparity estimation in an unsupervised manner. Our method performs better than dictionary based methods. In future, we can add explicit regularization [9] to improve the disparity maps. The hardware realization of this work using the same setup of Marwah et al. Also, we can look at light field generation from a single RGB image.

References

[1] Lytro. https://www.lytro.com/imaging, 2017 (accessed June 20, 2017). 1, 2

[2] Our project page. http://www.ee.iitm.ac.in/comp_photolab/project-clf-recovery.html, 2017 (accessed June 20, 2017). 1, 4, 5

[3] E. H. Adelson and J. Y. Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992. 1

[4] S. D. Babacan, R. Ansorge, M. Luessi, P. R. Matarán, R. Molina, and A. K. Katsaggelos. Compressive light field sensing. *IEEE Transactions on image processing*, 21(12):4746–4757, 2012. 2

[5] T. E. Bishop, S. Zanetti, and P. Favaro. Light field superresolution. In *Computational Photography (ICCP), 2009 IEEE International Conference on*, pages 1–9. IEEE, 2009. 2

[6] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016. 2, 3

[7] L. G. La photograhie integrale. In *Academic des Sciences*, pages 31–42. ACM, 1908. 2

[8] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala. Spatio-angular resolution tradeoffs in integral photography. *Rendering Techniques*, 2006:263–272, 2006. 1, 2

[9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *arXiv preprint arXiv:1609.03677*, 2016. 2, 3, 5, 7

[10] P. Imaging. Capture life in 3d. https://www.pelicanimaging.com/, 2017 (accessed July 10, 2017). 2

[11] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1547–1555, 2015. 4, 5

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 4

[13] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):193, 2016. 2, 3, 4, 5

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[15] A. Levin and F. Durand. Linear view synthesis using a dimensionality gap light field prior. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1831–1838. IEEE, 2010. 2

[16] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42. ACM, 1996. 1, 2

[17] J. Mairal, F. Bach, J. Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014. 6

[18] X.-J. Mao, C. Shen, and Y.-B. Yang. Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections. *arXiv preprint*, 2016. 3

[19] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):46, 2013. 2, 6

[20] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 22–28. IEEE, 2012. 2

[21] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a handheld plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005. 1, 2

[22] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 765–776. ACM, 2005. 1

[23] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 24–32, 2015. 2