

# **BGDisp-ResNet: A Robust Disparity Estimation and View Synthesis Pipeline Integrating with Bilateral 3D Grid Features in Deep Residual Networks for Light Field Cameras**

*An M Tech project report*

*submitted by*

**SRICHARAN VEGGALAM**

*in partial fulfilment of the requirements  
for the award of the degree of*

**MASTER OF TECHNOLOGY**

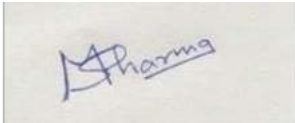


**DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.**

**JUNE 2021**

# THESIS CERTIFICATE

This is to certify that the thesis entitled **BGDisp-ResNet: A Robust Disparity Estimation and View Synthesis Pipeline Integrating with Bilateral 3D Grid Features in Deep Residual Networks for Light Field Cameras**, submitted by **Sricharan Veggalam (EE19M072)**, to the Indian Institute of Technology Madras, in partial fulfillment of the requirements for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under my 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. Mansi Sharma**  
Research Guide  
INSPIRE FACULTY  
Dept. of Electrical Engineering  
IIT-Madras, 600 036

Place: Chennai

Date: 18th June 202

## **ACKNOWLEDGEMENTS**

I would like to thank Dr. Mansi Sharma for guiding me and extending me her valuable assistance throughout the course of this M Tech project.

# ABSTRACT

**KEYWORDS:** Light Field; View Synthesis; Bilateral Grid; Disparity Estimation

Light field imaging has become widespread recently with the introduction of consumer light field cameras. However, these light field cameras often sparsely sample in either spatial or angular domain due to an inherent trade-off between the angular and spatial resolution. In this thesis, to mitigate this trade-off an end-to-end deep learning framework was used which specifically takes only a sparse set of input views to synthesize new set of views.

Building upon existing view synthesis techniques this process has been broken down into two parts i.e. disparity estimation and color estimation components. A deep residual neural network followed by convolutional neural network was used to model the two components and the whole learning framework can be trained end-to-end by minimizing the error between synthesized and ground truth images.

The proposed approach uses only four corner sub-aperture views from the light fields captured from light field cameras. The input feature map to the network is computed using 3D bilateral grids obtained from the four corner views which enables edge-aware processing of light field images. The experimental results show that the proposed approach synthesizes high-quality images compared to other techniques on a variety of real world scenes. This approach could potentially decrease the required angular resolution of consumer light field cameras, which allows their spatial resolution to increase.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>i</b>
<b>ABSTRACT</b>	<b>ii</b>
<b>LIST OF FIGURES</b>	<b>iv</b>
<b>ABBREVIATIONS</b>	<b>v</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Problem definition . . . . .	2
1.3 Summary of our work . . . . .	3
1.3.1 Organization of our thesis . . . . .	4
<b>2 RELATED WORK</b>	<b>5</b>
2.1 Depth-dependent view synthesis . . . . .	5
2.2 Depth-independent view synthesis . . . . .	6
<b>3 OVERVIEW OF KEY CONCEPTS</b>	<b>8</b>
3.1 Bilateral Grid . . . . .	8
3.1.1 Construction of Bilateral Grid . . . . .	8
3.1.2 Why 3D bilateral grids . . . . .	9
3.2 Convolution residual networks (ResNets) . . . . .	10
3.2.1 Convolution neural networks (CNNs) . . . . .	10
3.2.2 ResNets . . . . .	11
<b>4 PROPOSED ALGORITHM</b>	<b>14</b>
4.1 Disparity estimator . . . . .	14
4.2 Color predictor . . . . .	16
4.3 Results . . . . .	17
4.4 Conclusion . . . . .	19

## LIST OF FIGURES

1.1	Using only four corner sub-aperture images of a light field with angular resolution $8 \times 8$ , all other views of the light field are synthesized by our proposed approach. . . . .	2
3.1	Construction of bilateral grid . . . . .	9
3.2	A CNN sequence for handwritten digits classification . . . . .	10
3.3	Training and testing error% for 20-layer and 56-layer CNN . . . . .	11
3.4	Building block for residual learning . . . . .	12
3.5	Comparison of error% between ResNets and plain CNNs . . . . .	13
4.1	Pipeline of our system. Architectures of the networks used for disparity estimator and color predictor components are shown in fig.4.2 and fig.4.3 respectively. . . . .	14
4.2	Disparity Network used for estimation of disparity. This network is based on the ResNet-18 architecture. We newly added a $1 \times 1$ convolutional layer at the end of each block as shown in figure. . . . .	15
4.3	Color CNN consists of above convolutional layers each followed by rectified linear unit (ReLU). . . . .	16
4.4	Synthesized novel views at $q = (5,5)$ coordinate are shown above and are compared with Ground truth (GT). In the estimated disparity darker pixels indicate the regions that are closer to the camera. Our algorithm produces reasonable disparities for the purpose of view synthesis. . . . .	17
4.5	Our approach compared against other methods. . . . .	18
4.6	Our approach compared against other methods. . . . .	19

## ABBREVIATIONS

<b>IITM</b>	Indian Institute of Technology, Madras
<b>AR</b>	Augmented Reality
<b>VR</b>	Virtual Reality
<b>DL</b>	Deep Learning
<b>CNN</b>	Convolutional Neural Network
<b>BG</b>	Bilateral Grid
<b>ResNet</b>	Residual Network
<b>ReLU</b>	Rectified Linear Unit
<b>CV</b>	Computer Vision
<b>PSNR</b>	Pixel Signal-to-Noise Ratio
<b>SSIM</b>	Structural Similarity Index
<b>GT</b>	Ground Truth

# CHAPTER 1

## INTRODUCTION

### 1.1 Background and Motivation

Light Field imaging as a revolutionary imaging technology, has attracted extensive attention from both academia and industry, especially with the emergence of commercial plenoptic cameras and recent dedication in the field of Virtual Reality and Augmented Reality. Equipped with additional optical components like the microlens array inserted between the main lens and the image sensor, plenoptic cameras are capable of capturing both intensity and direction information of rays from real-world scenes, which enables applications such as refocusing and 3D display. However, due to the limited sensor resolution an inherent tradeoff between angular and spatial resolution inevitably occurs, which restricts light field imaging in many practical vision applications. View synthesis, which synthesizes novel views from a sparse set of input views captured using light field cameras is one of the possible solution to this problem.

Generally, existing traditional view synthesis approaches [Chaurasia *et al.*; Wanner and Goldluecke] typically first estimate the depth at the input views and use it to warp the input images to the novel view. They then combine these images in a specific way (for example by weighting each warped image) to obtain the image of the novel view. Building upon these methods we break down the task of view synthesis into disparity (depth) estimation and color predictor components. In our approach we use two networks connected sequentially for estimating the disparity and pixel colors of the novel view, respectively. Since both the networks are trained simultaneously by minimizing the error between the synthesized novel view and the ground truth, the required disparity for view-warping is implicitly produced by the first network, which is more suitable for view synthesis application. Bilateral grids obtained from four corner images were used to compute the input feature map, this enables the system to learn weights which will be capable of producing edge-consistent novel views.



## 1.2 Problem definition

Formally, the problem of novel view synthesis can be defined as follows:

Given the position  $q$  of the novel view and a sparse set of input views  $L_{p_1}, L_{p_2} \dots, L_{p_N}$  the goal is to synthesize the image  $L_q$  at the novel view. We can express this as follows:

$$L_q = f(L_{p_1}, L_{p_2}, \dots, L_{p_N}, q) \quad (1.1)$$

where  $p_i$  and  $q$  refer to the coordinates  $(x, y)$  of the input view and the novel view, respectively. Here, the relationship between the input views and the novel view is defined by the function  $f$ . As it requires finding connections between all the input views, and collecting appropriate information from each image based on the position of the novel view, this relationship defined by  $f$  is very complex. Inaccuracies such as noise and optical distortions in the light field images of real-world scenes further add to the complexity of this relationship. This relationship is learnt by our model which consists of a deep residual neural network followed by a convolutional neural network, through training.



Figure 1.1: Using only four corner sub-aperture images of a light field with angular resolution  $8 \times 8$ , all other views of the light field are synthesized by our proposed approach.

### 1.3 Summary of our work

The performance of our view synthesis pipeline is demonstrated using only the four corner sub-aperture views obtained from  $8 \times 8$  light fields captured using consumer light field cameras (see fig.1.1). Experimental results show that our method outperforms multiple traditional view synthesis techniques. Since our method employs residual neural network it is observed that our model achieves very good PSNR (Peak Signal-to-Noise Ratio) in considerably less number of iterations during training even though we are working with large amount light field images (close to 80 images).

We believe that our system potentially could be used to decrease the required angular resolution of light field cameras, which allows their spatial resolution to increase. Our method can be used on a subset of four angular views to synthesize the in between views which can increase the baseline of consumer light field cameras, this is another application of our approach. In summary, we make the following contributions:

We present a robust deep learning frame work for novel view synthesis using consumer light field cameras. Our system includes the disparity estimator which is modeled using a convolution residual network (ResNet) and this network is fed with bilateral 3D grid features which results in better disparity estimation suited for edge-consistent view synthesis. Also due to the inclusion of ResNet our model can be trained to achieve good PSNR values in less time compared other approaches [Kalantari *et al.*] The output of the disparity estimator network is connected to the color predictor network which is modeled using a convolutional neural network, synthesizes the final novel view.

We propose a end-to-end deep learning framework where both the networks are trained simultaneously by directly minimizing the error between the synthesized and ground truth images which alleviates the need for explicit depth information for view-warping since this information is implicitly estimated by the disparity (ResNet) network. So our model produces disparities which are suitable especially for the view synthesis application.

### **1.3.1 Organization of our thesis**

The thesis is organised in the following order:

In chapter 1, we introduce the method of novel view synthesis and explain how it can be used to mitigate the inherent trade-off between angular and spatial resolution of the consumer light field cameras. We define the problem of view synthesis precisely and also summarize our work in this thesis.

In chapter 2, a brief review of the traditional view synthesis techniques is given. Few state-of-the-art algorithms are also explained.

In chapter 3, some key concepts that are used in our approach are explained. Bilateral grids are explained in detail about how they are constructed and why it is used in our approach. Also Convolutional neural networks and ResNets are explained specifying the advantages of ResNets over CNNs.

In chapter 4, we propose our algorithm and network architecture and explain each component of our pipeline in detail. We show our results and compare them with the results of multiple other view synthesis techniques. we conclude our thesis by briefly talking about our proposed algorithm and discuss about our contribution.

## CHAPTER 2

### RELATED WORK

Several powerful methods for increasing the resolution in both angular [Levin and Durand; Shi *et al.*; Wanner and Goldluecke] and spatial [Bishop *et al.*; Cho *et al.*] domains have been proposed to solve the problem of the light field’s limited resolution. In this work we focus on the techniques that are designed for angular super-resolution. We first review the algorithms which explicitly use depth information for view synthesis and then explain the depth-independent view synthesis approaches.

#### 2.1 Depth-dependent view synthesis

Depth-dependent view synthesis approaches generally synthesize novel views of a scene in a two-step process, i.e. first estimating disparities of the input views and then warping those input views to novel views based on the estimated disparities, followed by combining the warped images in a specific way (e.g. weighted summation) to obtain the final novel views.

To reconstruct images at novel views from an input light field Wanner and Goldluecke proposed an optimization approach. They reconstruct novel views using the depth estimates at the input views, by minimizing an objective function which maximizes the quality of the final results. Their method produces reasonable results on dense light fields but for sparse input views, it produces results with tearing, ghosting, and other artifacts as shown in our results. This is mainly because of two reasons. Firstly, independent of the view synthesis process, they estimate the disparity at the input views as a preprocessing step. Even state-of-the-art light field disparity estimation techniques [Wang *et al.*; Jeon *et al.*] are not typically designed to maximize the quality of synthesized views, due to this they are not suitable for this application. Secondly, Wanner and Goldluecke’s method assumes that the images are captured under ideal conditions but in practice this is not true since the images from consumer light field cameras are usually noisy and suffer from optical distortions.

A phase-based approach is proposed by Zhang *et al.* to reconstruct the light fields from a micro-baseline stereo pair. However, since their approach is iterative, it is often slow since and prevents its usage in practice. Using convolutional neural networks (CNN) Yoon *et al.* perform spatial and angular super-resolution on light fields . However, their method is not able to synthesize views at arbitrary locations and can only increase the resolution by a factor of two. The patch-based synthesis method by Zhang *et al.* decomposes the disparity map into different layers and requires user interactions for various Light Field editing goals but has limited performance for view synthesis and cannot handle challenging scenes.

Some approaches [Eisemann *et al.*; Chaurasia *et al.*] typically use multi-view stereo algorithms to estimate depth and use this depth to warp and combine input images to the novel view. However, these are not suitable for light fields with a narrow baseline.

Flynn *et al.* proposed a deep learning method to perform view synthesis on a sequence of images with wide baselines. They first project the input images on multiple depth planes, then estimate the pixel color and weight of the image at each depth plane from these projected images. Following that they compute a weighted average of the estimated pixel colors to obtain the final pixel color. However, their method is slow compared to our (few minutes vs. seconds).

## 2.2 Depth-independent view synthesis

To reconstruct the full 4D light field from a 3D focal stack sequence Levin and Durand use a prior based on the dimensionality gap. To reconstruct a dense light field from a 1D set of view points, Shi *et al.* leverage sparsity in the continuous Fourier spectrum. Using multidimensional patches from a sparse set of input views Schedl *et al.* reconstruct a full light field . These methods are not able to synthesize novel views at arbitrary positions and also they require the input samples to be captured with a specific pattern.

Marwah *et al.* employs a dictionary-based approach to reconstruct light fields using a coded 2D projection . However, their method requires the light fields to be captured in a compressive way. Using a Gaussian mixture model Mitra and Veeraraghavan introduce a patch-based approach to model the light field patches . However, this method struggles on low-quality images taken with commercial light field cameras and is not

robust against noise.

Other approaches synthesize images without explicitly estimating the geometry. For example, to reconstruct the image at a novel view Mahajan *et al.* propose to move the gradients in the input images along a specific path. A patch-based optimization framework is proposed by Shechtman *et al.* to reconstruct images at novel views. However, since these approaches work on only two input images they are not able to utilize all the information available in light fields.

View synthesis is model as learning-based angular detail restoration on 2D Epipolar Plane Images (EPIs) by Wu *et al.*. A “blur-restoration-deblur” framework is proposed that consists of following steps: firstly, using a predefined blur kernel the input EPI is convolved; secondly, to restore the angular detail of the EPI damaged by the under-sampling, a CNN is applied; finally, to recover the spatial detail suppressed by the EPI blur a non-blind deconvolution operation is applied. This method achieves promising results on a variety of scenes. However, the operations of “blur-restoration-deblur” loop numerous times before all the in-between views are synthesized and also the potential of the full LF data is underused. The important insight of view synthesis is to make full use of the input views. It is necessary that the input views are regularly spaced on a grid to reduce the difficulty of collecting data,

There are also other several algorithms that have approached this problem using deep learning. Dosovitskiy *et al.* trained a CNN to render images of chairs given a graphics code containing the rendering details. Expanding on this work, Yang *et al.* decode the implicit rendering information from the input image instead of representing it explicitly as the graphics code. Then the desired transformation is applied and the new view is rendered. To estimate appearance flow, Zhou *et al.* train a CNN and this flow is then used to warp the input image to the novel view. Since these methods are specifically designed to work on objects they do not work well on general scenes and moreover they only use a single image, and are unable to utilize all the images in light fields.

## CHAPTER 3

### OVERVIEW OF KEY CONCEPTS

#### 3.1 Bilateral Grid

For a variety of image enhancement and editing or manipulation techniques nonlinear filters have proved to be very useful since they take image structure into account. The bilateral filter in particular is also a nonlinear filtering process that respects strong edges while smoothing an image. Bilateral filtering has been widely used in computational photography and image processing applications but a common drawback of this method is the computational complexity for processing high-definition (HD) content.

Paris and Durand build upon the fast bilateral filter presented by Durand and Dorsey and they recast the bilateral filtering as a higher-dimensional space linear convolution followed by trilinear interpolation and a division. Chen *et al.* generalize the ideas presented by Paris and Durand and introduce a new higher dimensional compact data structure i.e. the bilateral grid which enables a number of edge-aware image manipulations on high resolution images in real time. The bilateral grid is a 3D representation of a 2D image. It separates the pixels of the image not only by spatial coordinate, but also by respective intensity value or range coordinate.

##### 3.1.1 Construction of Bilateral Grid

The first two dimensions  $(x, y)$  of the 3D bilateral grid correspond to 2D position in the image (gray-scale) plane and form the spatial domain, while the third dimension  $z$  corresponds to the image intensity.

Let  $I(x, y) = z$  be a gray scale image normalized to  $[0,1]$  where  $x, y$  are the pixel indices and  $z$  is the intensity value, its corresponding bilateral grid BG is given by

$$BG([x/s_s], [y/s_s], [z/s_i]) = z \quad \forall x, y, z \in I \quad (3.1)$$

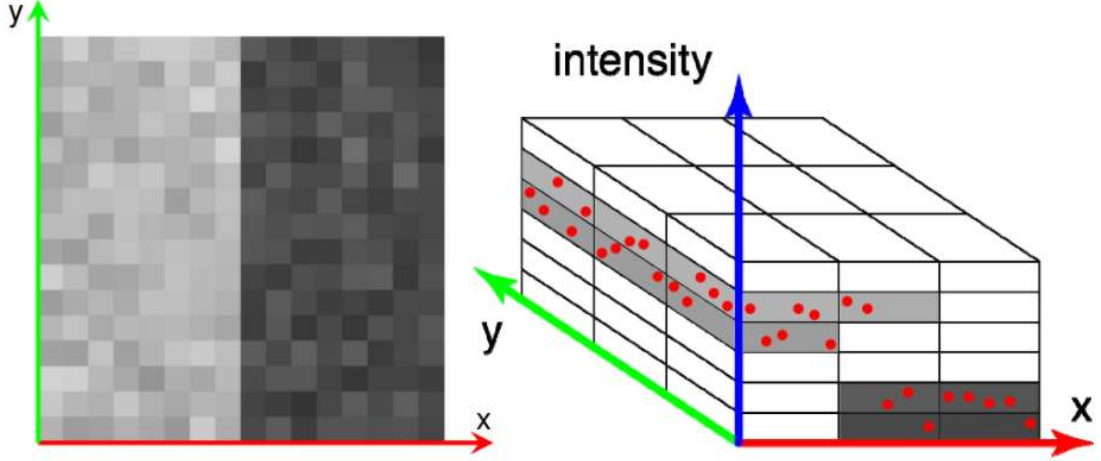


Figure 3.1: Construction of bilateral grid

Here,  $s_s$  and  $s_i$  are sampling rates in spatial axis and intensity axis, respectively and  $[\cdot]$  is the closest-integer operator. For images which are large adding a third dimension can make the size of 3D bilateral grid very large. To control the size of the grid above sampling rates can be used and their values are determined by the operation we want to perform on the grid. Intuitively,  $s_i$  controls the degree of edge preservation, while  $s_s$  controls the amount of smoothing. A smaller  $s_i$  or  $s_s$  yields a larger number of grid cells and requires more memory. In fig.3.1 the 3D array is constructed using the gray-scale image on the right in which we can see that the edge is separated by the intensity dimension as well.

### 3.1.2 Why 3D bilateral grids

We do not perform the processing inside the grid in our proposed framework but we exploit the edge-aware properties of the bilateral grid and enable our model to learn on high-dimensional feature space for edge consistent disparity estimation and novel view synthesis. The key point to note is that in the spatial dimension of an image, although two pixels across an edge are close, but from the bilateral grid perspective, they are distant from each other because their values differ widely in the intensity dimension of the data structure. Due to this property while performing convolution operations only a limited number of intensity values will get affected around edge pixels.



## 3.2 Convolution residual networks (ResNets)

### 3.2.1 Convolution neural networks (CNNs)

CNNs were first introduced by Yann LeCun in the 1980s. The early version of CNNs were able to recognize handwritten digits. CNNs found a good market in banking and postal services, where they were used to read digits on checks and zip codes on envelopes. At the time, the technique was only applicable to images with low resolutions because CNNs needed a lot of data and compute resources to work efficiently for large images. Now due to availability of large sets of data (eg: ImageNet), and vast compute resources enabled researchers to create complex CNNs which can perform computer vision tasks that were previously impossible.

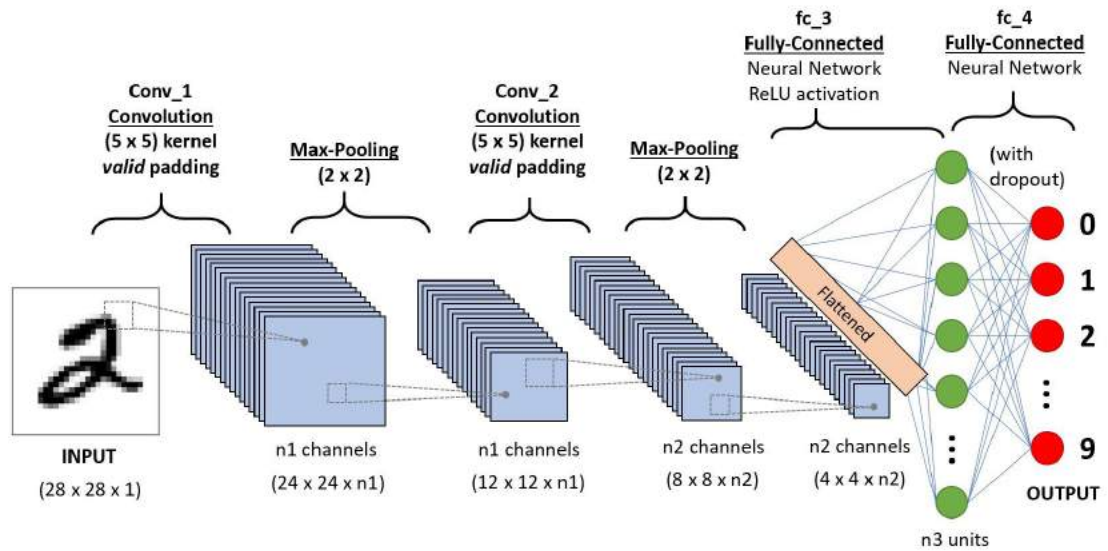


Figure 3.2: A CNN sequence for handwritten digits classification

In fig. 3.2 a CNN sequence which consists of convolutional, max-pooling and fully-connected which is used to classify handwritten digits is shown. Through the application of relevant filters a CNN is able to successfully capture the Spatial and Temporal dependencies in an image. Due to the reduction in the number of parameters involved and reusability of weights, the architecture performs a better fitting to the image dataset compared to the conventional feed-forward neural networks.

### 3.2.2 ResNets

We are getting state of the art results on problems such as image classification and image recognition with the introduction of deep CNNs. Due to this over the years, researchers having been adding more layers to make deeper neural networks hoping to solve such complex tasks and to also improve the classification/recognition accuracy. But, it has been observed that it becomes difficult to train as we go adding on more layers to the neural network, and also the accuracy starts saturating and then degrades also. This is where ResNets come into rescue and help us solve this problem.

Residual Network (ResNet) is a specific type of neural network that was introduced by Kaiming He et al. The intuition behind adding more layers to solve a complex problem is that these layers progressively learn more complex features. For example, in case of image recognition, the first layer may learn to identify edges, the second layer may learn to detect textures and similarly the third layer can learn to detect objects and so on. But it has been observed that there is a threshold for maximum depth in the context of traditional CNN model. In fig.3.3 the plot shows the error percent on training and testing data for a 20 layer and 56 layer CNN.

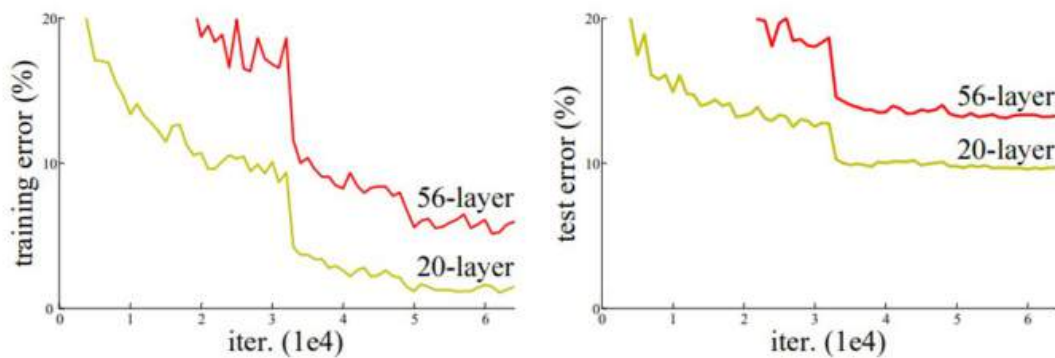


Figure 3.3: Training and testing error% for 20-layer and 56-layer CNN

We can see that error% for 56-layer is actually more than a 20-layer network for both training data as well as testing data. This shows that adding more layers to a network results in degradation of its performance. This could not be the result of overfitting because here the error% of the 56-layer network is worst on both training as well as testing data.

## Residual block

With the introduction of ResNet or residual networks the problem of training very deep networks has been solved and Residual Blocks are the building blocks of these Resnets.

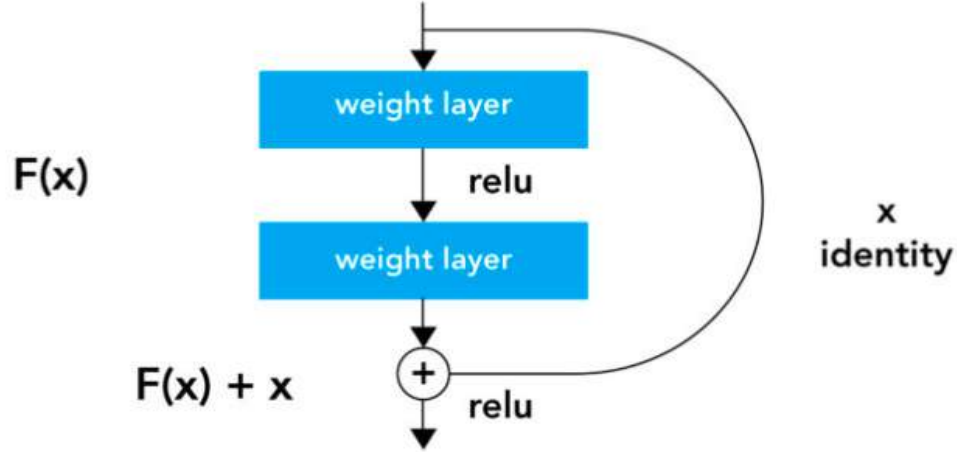


Figure 3.4: Building block for residual learning

Building block is ResNets is shown in fig.3.4. Here, there is a direct connection which skips some layers(may vary in different models) in between which is called 'skip connection' and is the core of residual blocks. The output of the layer changes due to this skip connection.

Without using this skip connection, the input  $x$  just gets multiplied by the weights of the layer and a bias term is added. Then it goes through activation function  $F$  and we get output  $H(x)$  as

$$H(x) = F(w.x + b) \quad \text{or} \quad H(x) = F(x) \quad (3.2)$$

The changes as below with the introduction of skip connection

$$H(x) = F(x) + x \quad (3.3)$$

A small problem with this approach is when the dimensions of the input vary from that of the output which can happen due to convolutional and pooling layers. When dimensions of  $x$  and  $F(x)$  are different, to match the dimension the projection method is used which is done by  $1 \times 1$  or other convolutional layer to input. In that case, the

output is

$$H(x) = F(x) + w1.x \quad (3.4)$$

where  $w1$  represents the additional weights added to match the dimensions.

### How ResNet helps

While training a deep neural network the vanishing gradients problem is an unstable behavior that we encounter. In this situation a deep multi-layer feed-forward network is unable to propagate useful gradient information for adjusting the weights from the output end of the model back to the layers near the input end of the model. The skip connections in ResNet acts as an alternate shortcut path for the gradient to flow through towards the input end of the model and solve the problem of vanishing gradient in deep neural networks. These skip connections also help by allowing the model to learn the identity functions which ensures that the higher layer will perform at least as good as the lower layer, and not worse.

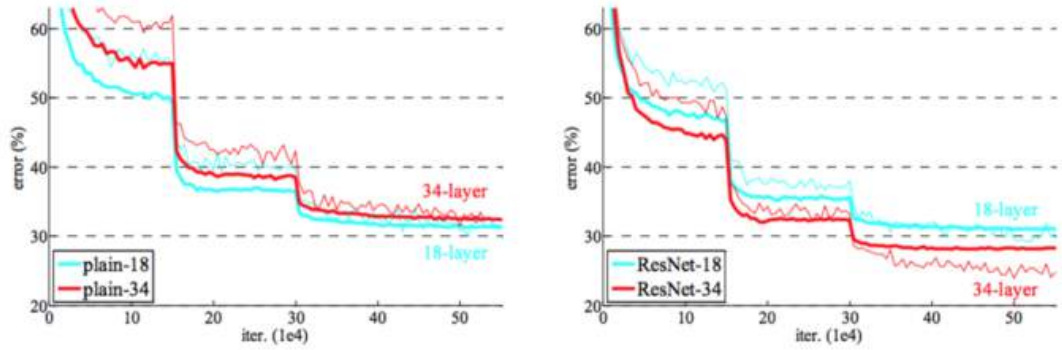


Figure 3.5: Comparison of error% between ResNets and plain CNNs

In fig.3.5, the plot shows that there is big difference in the networks with 34 layers. Here, ResNet-34 has much lower error% as compared to plain-34 CNN. This way ResNets enhance the performance of very deep neural networks.

# CHAPTER 4

## PROPOSED ALGORITHM

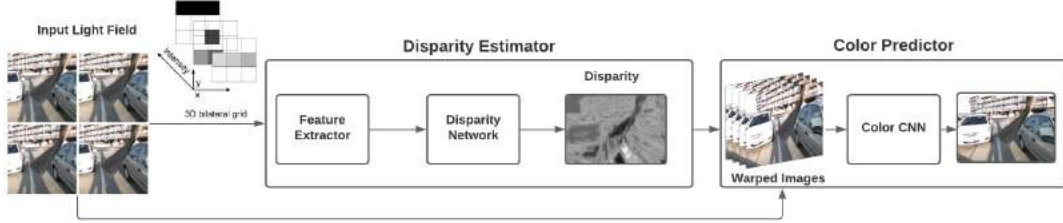


Figure 4.1: Pipeline of our system. Architectures of the networks used for disparity estimator and color predictor components are shown in fig.4.2 and fig.4.3 respectively.

### 4.1 Disparity estimator

The disparity at the novel view  $D_q$  is computed using disparity estimator component.

$$D_q = g_d(K) \quad (4.1)$$

where  $K$  represents the set of input features computed using the bilateral grids of the four input corner views and the relationship  $g_d$  modelled by the convolutional residual network (ResNet). We first convert the four corner RGB images to gray scale and then lift those gray scale images to 3D bilateral grids using the equation 3.1, here  $s_i$  (sampling rate of intensity axis) is chosen such that we get 10 intensity levels i.e. the length of the intensity dimension of the computed bilateral grid is 10. Now for every intensity channel of the four bilateral grids we perform warping operations using below equation

$$W_{p_i}^{d_l}(s) = L_{p_i}[s + (p_i - q).d_l] \quad (4.2)$$

where vector  $s$  contains the pixel position in  $x$  and  $y$  directions. Vectors  $p_i$  and  $q$  contains the position of input and novel views in  $x$  and  $y$  directions, respectively.  $L_{p_i}$

is the intensity channel of the  $i$ th input corner view. Here,  $W_{p_i}$  is the intensity channel obtained by warping  $L_{p_i}$  using the disparity level  $d_l$ . We consider  $L=10$  predefined disparity levels ( $l \in \{1, \dots, L\}$ ) in the range of  $[-21, 21]$  pixels. Then we compute mean and standard deviation of all  $W_{p_i}$ s at each disparity level as below

$$M^{d_l}(s) = \frac{1}{N} \sum_{i=1}^N W_{p_i}^{d_l}(s)$$

$$V^{d_l}(s) = \sqrt{\frac{1}{N} \sum_{i=1}^N (W_{p_i}^{d_l}(s) - M^{d_l}(s))^2}$$
(4.3)

Here  $N = 4$ , since we are using four corner views as input. The input feature map is then generated by concatenating the mean and standard deviation for all intensity and disparity levels. So,  $K = \{M^{d_1}, V^{d_1}, \dots, M^{d_L}, V^{d_L}\}$ . Since 10 intensity levels and 10 disparity levels are used, the feature vector is of 200 channels. This feature vector is fed to the disparity net shown in the fig.4.2. We avoid the alternative of training disparity estimator separately by minimizing the error between estimated and ground truth disparities because it requires ground truth disparities also which is difficult to obtain. Also, training the system end-to-end results in disparities better suited for novel view synthesis.

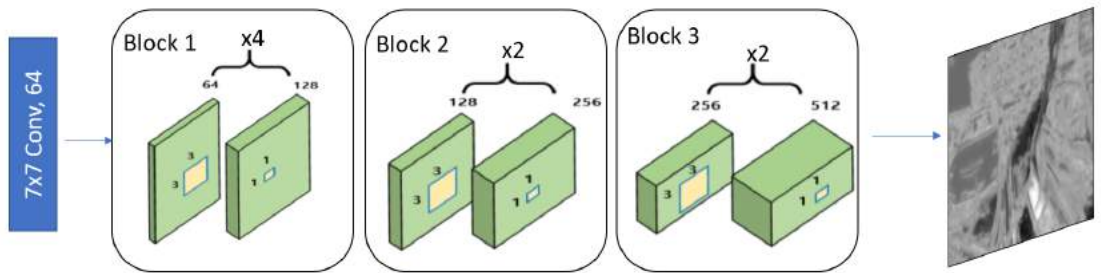


Figure 4.2: Disparity Network used for estimation of disparity. This network is based on the ResNet-18 architecture. We newly added a  $1 \times 1$  convolutional layer at the end of each block as shown in figure.

## 4.2 Color predictor

The final novel view ( $L_q$ ) is synthesized by using the a set of input feature that includes the warped images, the position of the novel view ( $q$ ) and the estimated disparity ( $D_q$ ):

$$L_q = g_c(H) \quad (4.4)$$

where relationship  $g_c$  is modelled using a CNN shown in fig.4.3. This relationship between the warped and final synthesized images is often complex because of occlusion, thus it is modelled using a CNN and learnt through training. The feature vector is  $H = \{W_{p_1}, \dots, W_{p_N}, D_q, q\}$  where

$$W_{p_i}(s) = L_{p_i}[s + (p_i - q) \cdot D_q(s)] \quad (4.5)$$

Here, the disparity gives useful information about the occlusion boundaries and the position of the novel view can potentially be used to weight a particular image more in synthesizing the novel view.

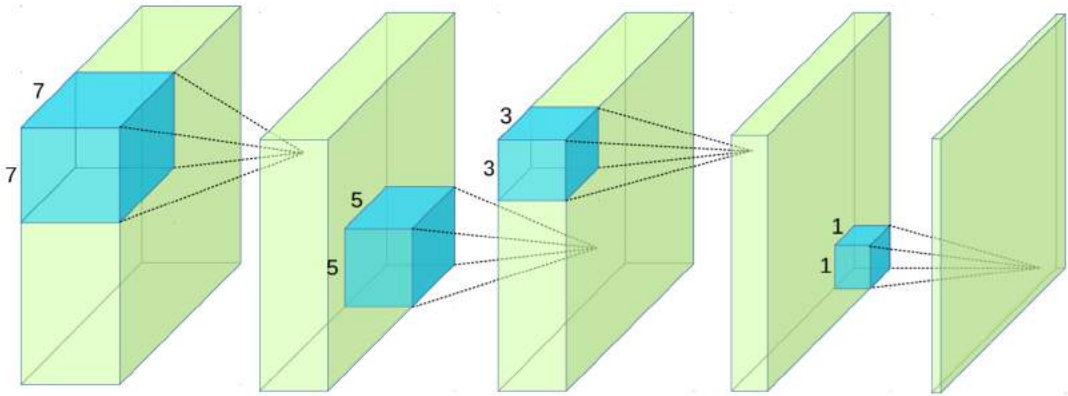


Figure 4.3: Color CNN consists of above convolutional layers each followed by rectified linear unit (ReLU).



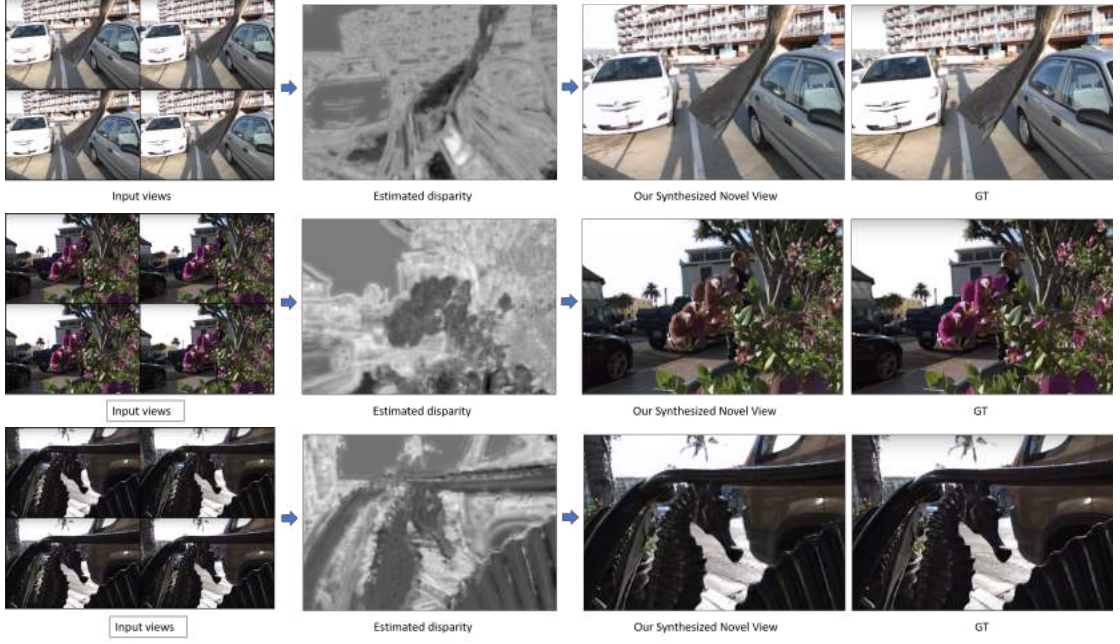


Figure 4.4: Synthesized novel views at  $q = (5,5)$  coordinate are shown above and are compared with Ground truth (GT). In the estimated disparity darker pixels indicate the regions that are closer to the camera. Our algorithm produces reasonable disparities for the purpose of view synthesis.

### 4.3 Results

The networks were trained by minimizing the  $L2$  distance (eqn 4.6) between the synthesized and ground truth images.

$$E = \sum_{k=1}^3 (\hat{L}_{q,k} - L_{q,k})^2 \quad (4.6)$$

where summation is over RGB channels,  $\hat{L}_{q,k}$  is the synthesized image at novel view and  $L_{q,k}$  is the ground truth image. Over 70 light field images captured with Lytro Illum camera were used as training set. To ensure diversity a variety of scenes with different lighting conditions, depth variations and texture properties were used. Patches of size  $60 \times 60$  with a stride of 16 pixels were extracted from the full images since training on the full images is slow. The output patches are then compared to the ground truth patches and the error at each pixel is back-propagated to adjust weights and train the networks.

The angular resolution of the light fields is  $8 \times 8$  from which only the four corner sub-aperture images were used as input to our system to generate the full light field. In



fig.1.1 we can see four corners views are used to synthesize the  $8 \times 8$  output grid. In comparative analysis we show one synthesized image where  $q=(5,5)$  for each scene and compare it with other techniques.

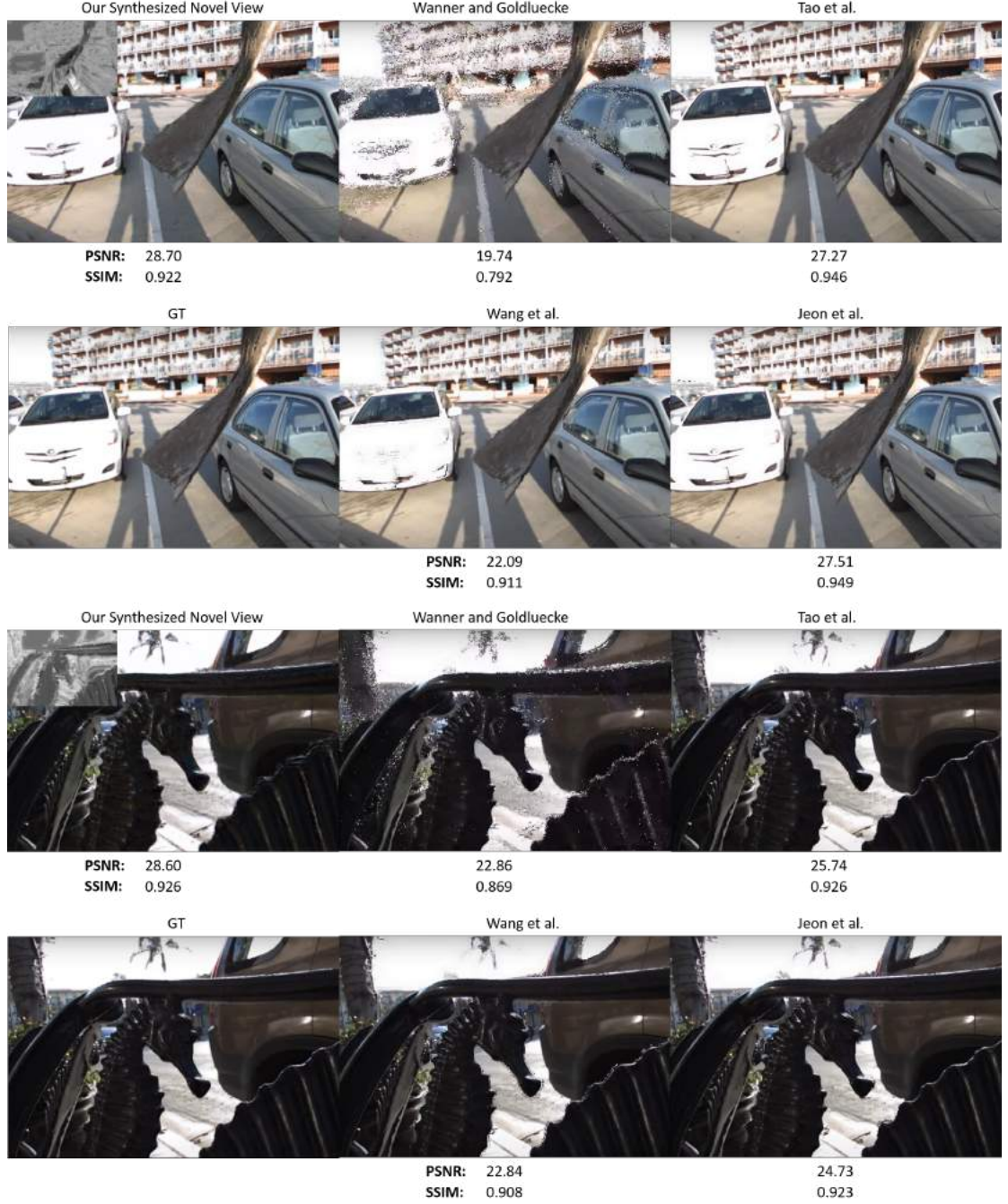


Figure 4.5: Our approach compared against other methods.

In Wanner and Goldluecke’s approach they compute the disparity for each input view using an existing technique first and then use those disparities in an optimization framework to the synthesize the novel views. In comparative analysis multiple light field disparity estimation techniques were adopted to generate the disparities required for Wanner and Goldluecke’s method. Specifically, Tao *et al.*, Wang *et al.*, Jeon *et al.*

techniques were used. Results were evaluated numerically, in terms of PSNR and structural similarity (SSIM) and are shown in the fig.4.5 and fig.4.6. SSIM of 1 indicates perfect perceptual quality with respect to the ground truth.



Figure 4.6: Our approach compared against other methods.

## 4.4 Conclusion

We have proposed a end-to-end learning-based approach for synthesizing novel views from a sparse set of input views captured with a consumer light field camera. Our system consists of disparity estimator which is modelled using ResNet and 3D bilateral grid features. And the color predictor component is modelled using a convolutional neural networks. The result of our approach were computed on a variety of light field scenes using only the four corner sub-aperture images captured with a Lytro Illum camera. Experimental results show that our method outperforms multiple other approaches.

## REFERENCES

1. **Bishop, T. E., S. Zanetti, and P. Favaro**, Light field superresolution. *In 2009 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2009.
2. **Chaurasia, G., O. Sorkine, and G. Drettakis**, Silhouette-aware warping for image-based rendering. *In Computer Graphics Forum*, volume 30. Wiley Online Library, 2011.
3. **Chen, J., S. Paris, and F. Durand** (2007). Real-time edge-aware image processing with the bilateral grid. *ACM Transactions on Graphics (TOG)*, **26**(3), 103–es.
4. **Cho, D., M. Lee, S. Kim, and Y.-W. Tai**, Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. *In Proceedings of the IEEE International Conference on Computer Vision*. 2013.
5. **Dosovitskiy, A., J. Tobias Springenberg, and T. Brox**, Learning to generate chairs with convolutional neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
6. **Eisemann, M., B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, C. Theobalt, and A. Sellent**, Floating textures. *In Computer graphics forum*, volume 27. Wiley Online Library, 2008.
7. **Flynn, J., 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*. 2016.
8. **Jeon, H.-G., 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*. 2015.
9. **Kalantari, N. K., T.-C. Wang, and R. Ramamoorthi** (2016). Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, **35**(6), 1–10.
10. **Levin, A. and F. Durand**, Linear view synthesis using a dimensionality gap light field prior. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.
11. **Mahajan, D., F.-C. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur** (2009). Moving gradients: a path-based method for plausible image interpolation. *ACM Transactions on Graphics (TOG)*, **28**(3), 1–11.
12. **Marwah, K., G. Wetzstein, Y. Bando, and R. Raskar** (2013). Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, **32**(4), 1–12.
13. **Mitra, K. and A. Veeraraghavan**, Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. *In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012.

14. **Schedl, D. C., C. Birklbauer, and O. Bimber**, Directional super-resolution by means of coded sampling and guided upsampling. *In 2015 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2015.
15. **Shechtman, E., A. Rav-Acha, M. Irani, and S. Seitz**, Regenerative morphing. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.
16. **Shi, L., H. Hassanieh, A. Davis, D. Katabi, and F. Durand** (2014). Light field reconstruction using sparsity in the continuous fourier domain. *ACM Transactions on Graphics (TOG)*, **34**(1), 1–13.
17. **Tao, M. W., S. Hadap, J. Malik, and R. Ramamoorthi**, Depth from combining defocus and correspondence using light-field cameras. *In Proceedings of the IEEE International Conference on Computer Vision*. 2013.
18. **Wang, T.-C., A. A. Efros, and R. Ramamoorthi**, Occlusion-aware depth estimation using light-field cameras. *In Proceedings of the IEEE International Conference on Computer Vision*. 2015.
19. **Wanner, S. and B. Goldluecke** (2013). Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, **36**(3), 606–619.
20. **Wu, G., M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu**, Light field reconstruction using deep convolutional network on epi. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
21. **Yang, J., S. E. Reed, M.-H. Yang, and H. Lee** (2015). Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. *Advances in neural information processing systems*, **28**, 1099–1107.
22. **Yoon, Y., 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*. 2015.
23. **Zhang, F.-L., J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu** (2016). Plenopatch: Patch-based plenoptic image manipulation. *IEEE transactions on visualization and computer graphics*, **23**(5), 1561–1573.
24. **Zhang, Z., Y. Liu, and Q. Dai**, Light field from micro-baseline image pair. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
25. **Zhou, T., S. Tulsiani, W. Sun, J. Malik, and A. A. Efros**, View synthesis by appearance flow. *In European conference on computer vision*. Springer, 2016.