

**DeepBGCRRFDisp-ResNet: An Integrated Deep Residual  
Network Based on Bilateral 3D Grid with Dense CRF  
model for Disparity Estimation and Refinement for High  
Quality View Synthesis**

*submitted by*

**BANOTHU NAVEEN KUMAR**

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

**MASTER OF TECHNOLOGY**

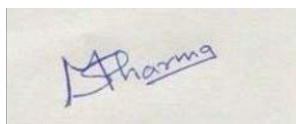


**DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.**

**JUNE 2020**

# THESIS CERTIFICATE

This is to certify that the thesis entitled **DeepBGCRCRFDISP-ResNet: An Integrated Deep Residual Network Based on Bilateral 3D Grid with Dense CRF model for Disparity Estimation and Refinement for High Quality View Synthesis**, submitted by **BANOTHU NAVEEN KUMAR (EE19M066)**, 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 bonafide 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: June 2020

## **ACKNOWLEDGEMENTS**

I would like to thank Dr. Mansi Sharma for the supervision and guidance throughout my M.Tech thesis study at IIT Madras.

# ABSTRACT

**KEYWORDS:** Novel View Synthesis ;Depth Estimation ;DenseCRFmodel ;Deep Learning ;convolutional neural network ;Computational Photography

Novel View Synthesis is a very important research problem in computer vision and computational photography which enables wide range of applications like re-cinematography, video enhancement, virtual reality etc..But,however developing these technologies faces several challenging technical problems which leads to occlusions and pixel inconsistency.So as a result it is challenging to align the input views together to synthesize the novel views from using those input images.

In this work,our framework first leverages a depth prediction model which estimates depth from the sparse set of input views which are captured using light field cameras. These depth maps are suitable for view synthesis tasks.

We then recover our depth structures that suffers from serious distortions near object counters based on coarse-to-fine dense CRF model.This methods recovers high quality depth maps from distorted ones with erroneous structures.

We develop a depth based view synthesis model by addressing identified requirements that are required while synthesizing like Geometrical consistent in-painting, Temporal consistency.We utilize context aware depth and color in-painting to fill in the missing information in the extreme views.

This thesis proposed novel view learning based pipeline to synthesize new views from a set of input views which integrates with deep residual neural network based on Bilateral 3D grid and dense CRF model for the efficient disparity estimation and refinement of high quality view synthesis.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>i</b>
<b>ABSTRACT</b>	<b>ii</b>
<b>LIST OF FIGURES</b>	<b>v</b>
<b>ABBREVIATIONS</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Problem definition . . . . .	3
1.3 Summary of our work . . . . .	3
1.4 Related work . . . . .	4
1.5 Organization of our thesis . . . . .	5
<b>2 Brief Overview Of Concepts</b>	<b>6</b>
2.1 Bilateral Grid . . . . .	6
2.1.1 computation of Bilateral Grid . . . . .	6
2.1.2 Construction of bilateral grids as color images . . . . .	7
2.1.3 How Bilateral Grids preserve edges and why Bilateral Grids	8
2.2 Convolutional Neural Network . . . . .	8
2.2.1 Convolutional Layer . . . . .	9
2.2.2 Pooling Layer . . . . .	10
2.2.3 Fully Connected Layer . . . . .	11
2.3 Residual Network . . . . .	11
2.4 Dense CRF model . . . . .	13
2.4.1 Related Work . . . . .	14
<b>3 Disparity Estimator</b>	<b>15</b>
3.1 Motivation . . . . .	15

3.2	Proposed network for disparity estimator . . . . .	15
3.2.1	Depth Structure Recovery . . . . .	16
<b>4</b>	<b>Novel View Synthesis</b>	<b>19</b>
4.0.1	Depth and color in-painting . . . . .	20
4.0.2	Context Extraction . . . . .	20
4.0.3	Point Cloud Rendering . . . . .	21
4.0.4	Summary . . . . .	21
4.1	Comparative Analysis . . . . .	21
4.2	Conclusion . . . . .	22
4.3	Future Work . . . . .	22

# LIST OF FIGURES

1.1	By using only four sub aperture images as our input we capture light field with angular resolution of 8*8 using lytro camera. Here our learning based method is able to handle occulsions between the rock and the background and produce a superficially fair image which is comparable to the GT image . . . . .	2
2.1	2D image represented as 3D BL grid . . . . .	6
2.2	Illustration of BL grid preserve edges . . . . .	8
2.3	Convolutional Neural Network . . . . .	9
2.4	Different Low level and high level features . . . . .	10
2.5	Max Pooling . . . . .	11
2.6	Vanishing Gradient Problem . . . . .	12
2.7	Residual Network . . . . .	12
3.1	Framework of our method . . . . .	15
3.2	Pipeline of our convolutional Residual Network . . . . .	16
3.3	(top)distorted depth map (bottom) Coarse-to-fine depth map . . . . .	17
4.1	Pipeline of our Novel View Synthesis Network . . . . .	19
4.2	Comparison of our approach against other methods . . . . .	21

## ABBREVIATIONS

<b>IITM</b>	Indian Institute of Technology, Madras
<b>ML</b>	Machine Learning
<b>DL</b>	Deep Learning
<b>CNN</b>	Convolutional Neural Network
<b>GT</b>	Ground Truth
<b>NVS</b>	Novel View Synthesis
<b>BG</b>	BiLateral Grid
<b>ResNet</b>	Residual Network
<b>CRF</b>	Conditional Random fields
<b>FCN</b>	Fully Convolutional Network
<b>ReLU</b>	Rectified Linear Unit
<b>BN</b>	Batch Normalization
<b>AI</b>	Artificial Intelligence
<b>CV</b>	Computer Vision
<b>NN</b>	Neural Network
<b>PSNR</b>	Peak Signal to Noise Ratio



# CHAPTER 1

## Introduction

### 1.1 Introduction

Novel View Synthesis(NVS) targets at generating novel views points of a scene or an object given only one image or few images of it.Light fields provide a quality representation of the real world scenes by enabling exciting applications such as viewpoint change and refocusing.Early light field cameras are bulky,expensive and thus not available to the general public and they also need a custom made setup.

Inspired by the success in deep learning recently in a wide variety of applications such as super resolution,deblurring and image denoising etc...So we propose to use a CNN to first build a depth prediction model which estimates depth from sparse set of inputs which helps for the view synthesis tasks.

Existing view synthesis based approaches first estimate depth maps at the input views and it wrap the input views to get the novel view and then they combine these images in a specific way.Here in this method,From the input image and its associated depth map we synthesize a sequence of novel views to produce the output while addressing specific requirements like Temporal Consistency,Geometrical consistent in-painting and we utilize context aware depth and colour in-painting to fill in the missing information in the extreme areas which are caused due to dis-occlusion.

Key contributions of this thesis are as follows:We introduce the problem of novel view synthesis from the set of input views which integrates with residual network on bilateral grid and dense CRF model for efficient depth estimation and high quality view synthesis.Experiments on real world imaginary demonstrates the effectiveness of our model.Our study shows that our system enables users to achieve better results when compares to all the existing methods for novel view synthesis.



(a) Input Views



(b) (top)Ours (bottom)reference



(c) Our Reconstructed Novel View

Figure 1.1: By using only four sub aperture images as our input we capture light field with angular resolution of  $8 \times 8$  using lytro camera. Here our learning based method is able to handle occlusions between the rock and the background and produce a superficially fair image which is comparable to the GT image

## 1.2 Problem definition

Given a set of input views say  $L_{p1}, L_{p2}, \dots, L_{pN}$  and the position of novel view say  $K$ . So, our goal is to estimate the image at the novel view  $L_K$ . We can formally write this as:

$$L_K = f(L_{p1}, \dots, L_{pN}, K) \quad (1.1)$$

Here,  $f$  is a function which defines the relation between novel view and input views.  $p_i$  and  $K$  refers to the  $(u, v)$  coordinates of input and novel view respectively. Inaccuracies in the form of optical distortions and noise will further more add to the complexity of this relationship. So we propose to use CNN as our learning model. So we model the function  $f$  with a CNN, In this case it takes the input views as well as the position of the novel view and outputs the image at the novel view. Since the relationship is complex and the network needs to find the connections between distant pixels, this solution produces blurry results and makes the training difficult.

We make the training more easy to control by following pipeline of the existing novel view synthesis methods and break our network into disparity and novel view synthesis components. Our contribution is to model each component and train both the model simultaneously by minimizing the error.

## 1.3 Summary of our work

We Perform our approach only on four sub-aperture views from 8\*8 views captured from Lytro Illum camera. Results show that our approach outperforms state-of-art schemes on challenging cases. Our method is two orders of magnitude faster than recent learning based deep stereo methods which takes only around 15 seconds to synthesize an image. We make the following contributions in this thesis:

- Our model consists of disparity and Novel view synthesis components which we model using two sequential CNN components.
- We integrate Bilateral 3D Grid features and Dense CRF model in Deep Residual Network to perform edge aware computations and for efficient Disparity Estimation.

- We train this network by minimizing the error between synthesized and GT images.
- Since we train our network in this way, Disparities are suitable for the view synthesis application.

## 1.4 Related work

The light field limited resolution problem has been extensively studied in the past and several methods for Novel View Synthesis which is based on depth maps have been proposed. We now review some algorithms that specifically work on light field images and explain the approaches for general scenes that perform view synthesis.

Novel View Synthesis-Variational Light Field Analysis for Disparity and super resolution Wanner and Goldluecke (2013) proposed an approach to reconstruct the images at novel views from a light field image. The given depth estimates at the input views reconstruct the novel views by minimizing with an objective function which in turn maximizes the quality of the final results. This method produces results on dense light fields but for sparse input views, it produces results with ghosting, tearing. One of the reasons for such result is that this method assumes that the images are captured under ideal conditions but in practice however the images light field cameras are noisy and suffer from distortions and the other reason being this method estimates disparity at the input views as a preprocess independent of the view synthesis process.

View synthesis for a scene-One category of approach Eisemann *et al.* (2008) proposed synthesis of novel views of a scene in a two step process. This method first estimate the depth from the input views and use the depth to warp input images to the novel view. They produce final views by combining these warped images. Unlike these approach, here in our method we use machine learning to model depth and colour components.

View synthesis from a single image-Recent novel view synthesis methods use single image setting Tatarchenko *et al.* (2015) Synthesizing novel views from a single image is quite challenging and they are often applicable to some specific scene types. But in this method although we use sparse input views, we focus on predicting depth maps that are suitable for the high quality view synthesis. We improve the estimated depth

directly and thus the estimated scene geometry to suppress artifacts such as inaccurate depth boundaries. While we would like also perform depth based view synthesis, we focus on predicting the depth maps that are suitable for the high quality view synthesis. Especially, we directly improve the estimated depth and tailor the depth prediction to the task of view synthesis.

Single Image depth Estimation -Gained a lot of research interest over past decades Koch *et al.* (2018). But, However depth from single image remains an open research problem because the quality of predicted depth map varies depending on image type and depth maps from all the existing methods does not produce high quality view synthesis results and are mostly not suitable. So make use of sparse set of input views in this thesis and produce high quality depths that are suitable for view synthesis tasks.

## 1.5 Organization of our thesis

The thesis organized in the order in which this research is carried out.

- In chapter 1, we have discussed how novel view synthesis is a very important research problem in computer vision and computational photography. We have also summarized our work and also reviewed briefly of the work that has been done in the recent past on novel view synthesis in this chapter.
- In chapter 2, We discuss about the brief overview of concepts like bilateral 3D grid, Convolutional Neural Network, Dense CRF model etc..
- Chapter 3 proposes a convolutional Neural Network based depth estimator and a comparative analysis of estimated depths to the state-of-art technologies.
- chapter 4 proposes a Novel View synthesis based pipeline using the disparity estimated maps from first CNN and then we conclude our work by talking briefly about our proposed network and the discuss about our contribution. Finally we discuss about future scope that can be carried out by our research.

# CHAPTER 2

## Brief Overview Of Concepts

### 2.1 Bilateral Grid

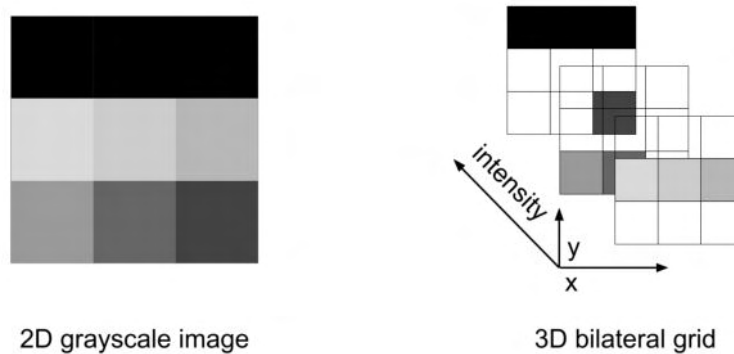


Figure 2.1: 2D image represented as 3D BL grid

Bilateral Grid is a new data structure introduced by Chen *et al.* (2007) that enables fast edge aware image processing. They preserve and retrieve edges present in the original image even after performing computations. Its a 3D representation of a 2D image which separates pixels not only by their spatial positions but also by their intensity values.

#### 2.1.1 computation of Bilateral Grid

Let the gray scale image  $I$  is represented as  $I(p,q)=r$  where 'r' is the intensity value at the pixel position  $(p,q)$ . Here  $p,q$  are the pixel indices values. The corresponding Bilateral grid is constructed as:

$$BG(p, q, r) = r \quad \forall p, q, r \in I \quad (2.1)$$

The motivation behind the fact that the network we trained is a 3D CNN is that any 2D operation on image space becomes a 3D operation in bilateral space.

Images can be quite large and adding a third dimension will even blow up the size of the image. We usually reduce the spatial resolution of the bilateral grid by a factor of

2 or 4 or sometimes even to higher factors if the image is of high resolution. Generally the choice of intensity dimension is either 32 or 64. On further experimentation it was easier for the networks to learn instead of  $BG(p,q,r)=r$ , we use

$$BG(p, q, r) = 1 \quad \forall p, q, r \in I \quad (2.2)$$

and thus we make gradient values equal for all the intensities during the back propagation and thus learning all the intensities at equal rates.

### 2.1.2 Construction of bilateral grids as color images

BL grid is generally constructed for grey scale images. But if we want to represent color images we would need a higher 5 dimensional grid with extra dimensions for RGB values. So to address this issue we would instead consider three BL grids separately one for each color channel. We consider each color image as a grey scale image. Let  $I$  be the image to be converted and let  $BG_r, BG_g, BG_b$  be the BL grids corresponding to each color channel. Let RGB be represented as

$$I(p, q) = (r, g, b) \quad (2.3)$$

$$BG_r(p, q, r) = 1, \quad BG_g(p, q, g) = 1, \quad BG_b(p, q, b) = 1 \quad (2.4)$$

$$BG_r(p, q, s) = r/s, \quad BG_g(p, q, s) = g/s, \quad BG_b(p, q, s) = b/s \quad (2.5)$$

$$\forall p, q, r, g, b \in I \text{ and where, } s = (r + g + b)$$

The above representation is proved to be the best from experiments and models were able to generalize to different scenes in images and learned faster.

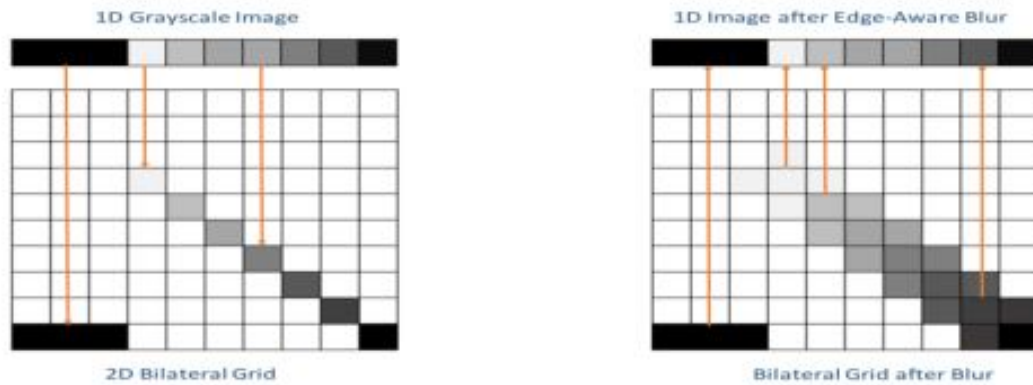


Figure 2.2: Illustration of BL grid preserve edges

### 2.1.3 How Bilateral Grids preserve edges and why Bilateral Grids

If we consider an example of blurring an image and when we use Gaussian kernel to blur, it affects all the neighbourhood pixels equally eventually destroying the sharp edges. Gaussian blur in  $N$  dimensions is treated as a Gaussian blur in  $N + 1$  dimensions when using bilateral grids. Thus, the 2D Gaussian blur kernel extends to 3D and so does the kernel. Since edges by definition mean sharp rise or fall in intensity values and since BL grids separate pixels by their intensity values too, the kernel never affects pixels on the other side of the edge. If we observe the figure above and notice how the first three black pixels are untouched and separation remains sharp even after blur operation is performed.

## 2.2 Convolutional Neural Network

CNN is a bizarre combination of biology, mathematics and little computer science involved. But this combo has been the most important influential innovation in the field of Computer vision(CV).CNNs are an integral part of the service of every company, small or big. They are the only reason why deep learning(DL) is so popular today. They can do things that we would have never imagined a computer would be capable of doing but nevertheless, they have their limitations and few fundamental drawbacks.

We see a significant improvement in challenging tasks like image recognition and classification since the concept of Deep CNN has been rolled out. Deeper networks will improve the performance of neural networks but the stacking of layer brought in a new



problem of vanishing gradients. This problem has been mitigated with the introduction of a new Residual Network (ResNet). This network introduces a skip connection between layers, the output of previous layers are added to the output of stacked layers in the feed-forward network. Doing this not only reduces the number of parameters but also helps in concatenating the feature maps for better gradient flow.

CNN is an algorithm which takes an input image assign learnable biases and weights to various objects that are present in the image and be able to differentiate one from the other. CNN network consists of input and output and several hidden layers in between. The hidden layers include a pooling layers, fully connected layers and normalization layers. Input and output layers are masked up by a activation layer called ReLu and followed by a fully convolutional layer.

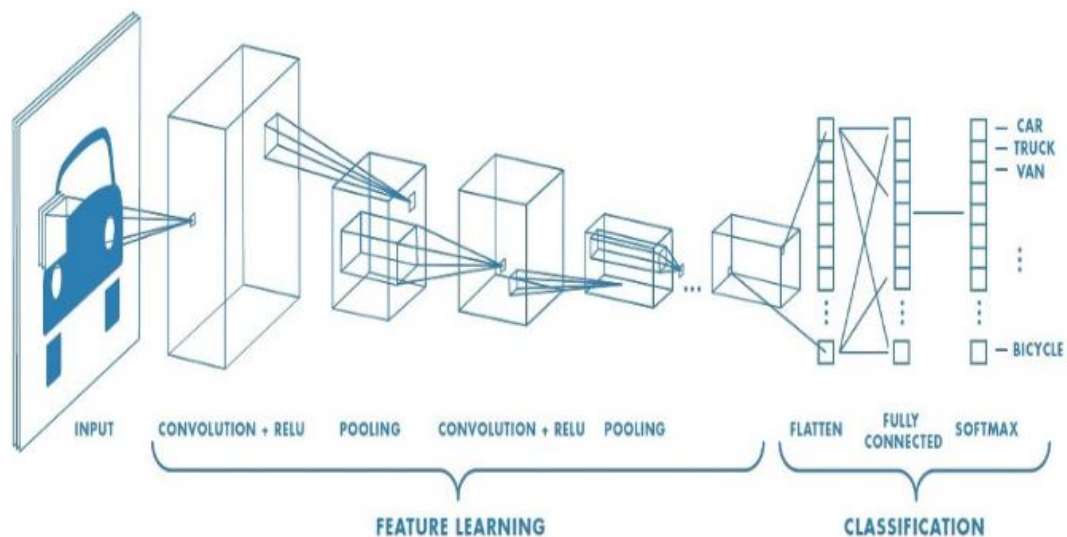


Figure 2.3: Convolutional Neural Network

This architecture of CNN is comparable to the structure of human brain and neural connectivity of cortex where the neurons respond to stimuli only in restricted region known as receptive field.

### 2.2.1 Convolutional Layer

The convolutional layer is the most essential block of CNN and it's objective is to extract high level features of the input image. Parameters of this image has some set of learnable filters which have a receptive field extended over the depth of the input image. First layer of convolution extracts the low level features such as color, edges etc...CNNs are not

restricted to one layer, with more layers more high level features are extracted which helps in better understanding of the images.

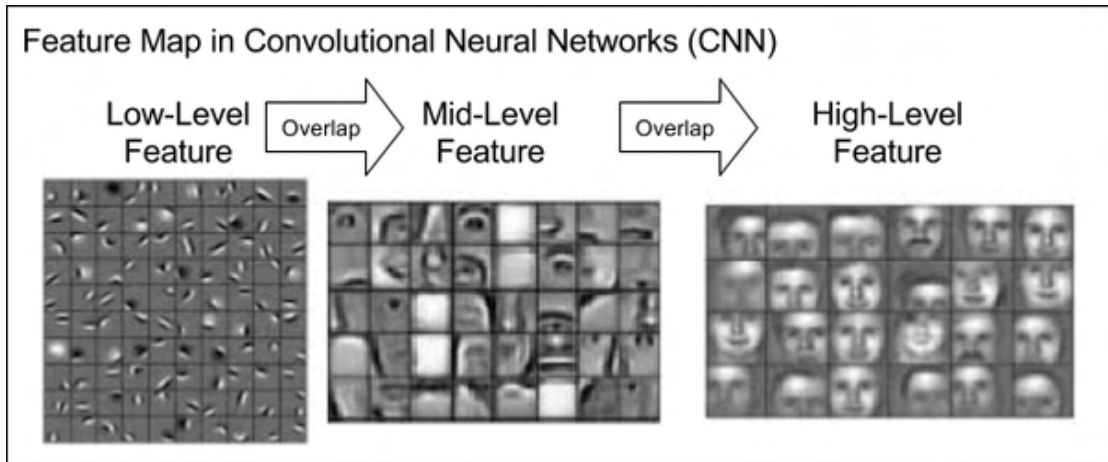


Figure 2.4: Different Low level and high level features

We have two types of convolution results. One is done by same padding and other by applying some valid padding. In same padding, the dimensionality of the image is either kept same or increased and in other padding output of convolution is reduced in dimensionality compared to the input image.

The formula to calculate output size is equals to:

$$O = (W - F + 2P) / S + 1 \quad (2.6)$$

Where O is the output height/length, W is the input height/length, F is the filter size, P is the padding and S is the stride.

### 2.2.2 Pooling Layer

Pooling layer is also responsible for reducing the dimension of the convolved feature like convolutional layer. It reduces the computational power required to process the data by reducing the dimensionality of images. There are two types of pooling. One is max pooling which returns the max value from the cluster of neurons and the other is average pooling which returns the average value from the cluster of neurons covered by kernel. Max pooling performs better than avg pooling because along with reducing the dimensionality it also suppresses the noise from the activations.

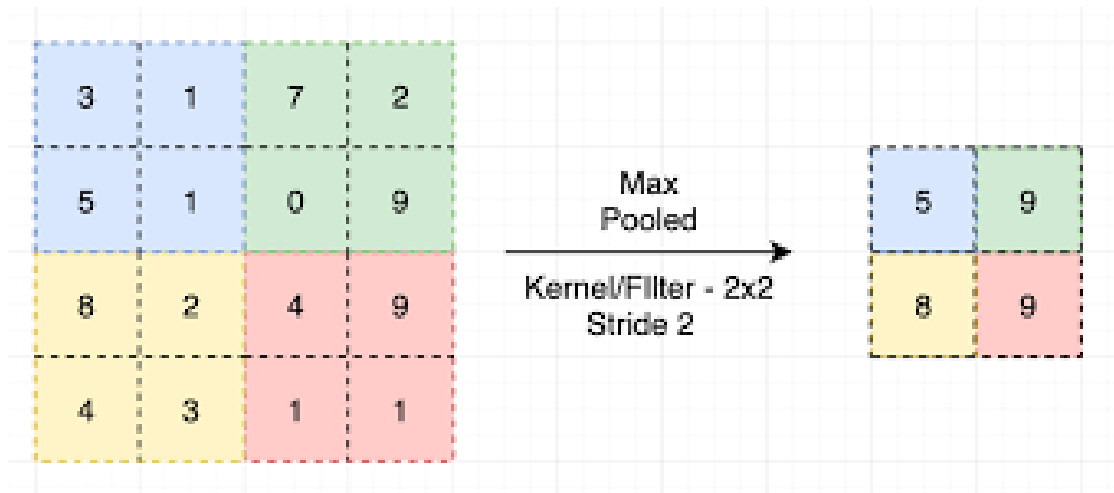


Figure 2.5: Max Pooling

### 2.2.3 Fully Connected Layer

The FC layer takes the input and gives the vector output whose dimension is equal to the no of classes. FC layer predicts the features of the class based on the output that is received from previous layer. It strongly correlates the high level features to a class in particular and calculates the probabilities of different other classes from the assigned weights.

## 2.3 Residual Network

Although accuracy increases with increase in the number of layers but there is a certain limit in the number of layers to be added to increase the accuracy Szegedy *et al.* (2015). If we keep on increasing the layers accuracy keeps increasing but at some point it becomes saturated. The networks faces dimensionality and vanishing gradient problems. Once it is saturated increasing layers would eventually degrade the accuracy. So this at this point shallower networks appear to be better than deeper networks and this problem is known as degradation problem.

Residual Networks (ResNets) He *et al.* (2016) have become popular in the field of Machine learning community in the last few years. They can train with hundreds of layers or even thousands without compromising on the performance of the network. Since then, there is a major change in the performance of applications such as object detection, Image classification and recognition.

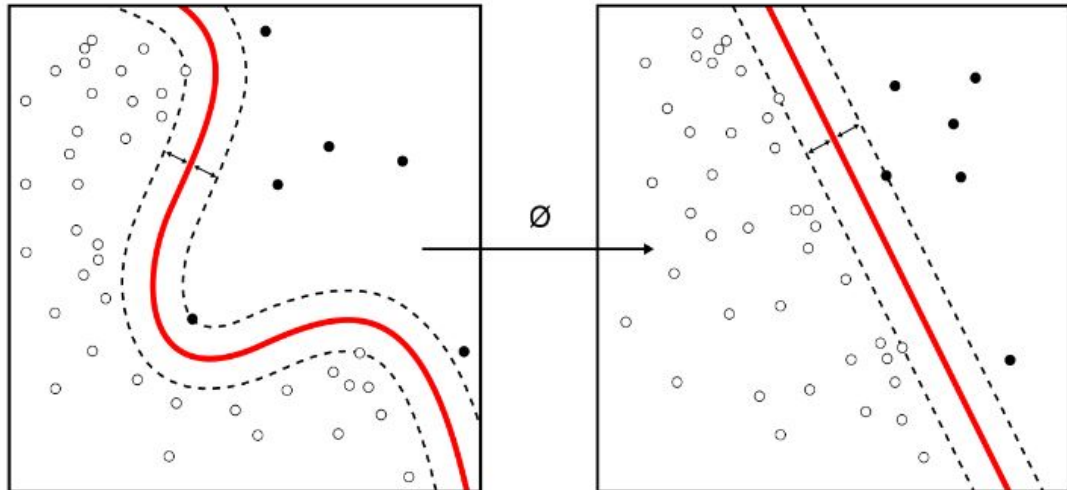


Figure 2.6: Vanishing Gradient Problem

Concept of Residual Networks is easy to understand. In normal conventional neural networks, Output of every layer is fed as an input to the next layer but in case of residual blocks the output of each layer is fed to next layer and also into the layers that is 3-4 blocks away.

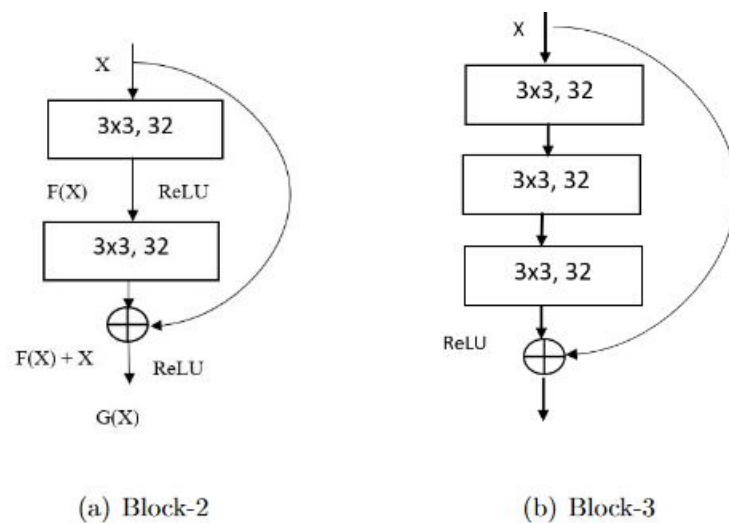


Figure 2.7: Residual Network

ResNets add skip connection between layers i.e., Output of previous layers are added to outputs of stacked layers in a feed forward manner. But as a matter of fact ResNets aren't the first to add this shortcut connections. Highway networks LeCun *et al.* (1998) have introduced gated shortcut connections. Highway networks can train networks with hundreds of layers effortlessly using gated units. These gates take care of the amount of information across the connections. But however, experiments show that Highway networks didn't outperform Residual Networks.

Giving a different perspective to Res Nets He *et al.* (2016) the authors proposed a pre-activation residual block Csáji *et al.* (2001) in which the gradient flow through shortcut connections to other layers without any interference. Shortcut connections are also known as identity connections since we can directly learn about identity function trusting on the shortcut connection only.

From fig 2.7(a), if we consider  $x$  to be the input and  $G(x)$  be the output distribution we can represent the difference to be:

$$G(x) = F(x) + x ; \quad F(x) = output - input = G(x) - x \quad (2.7)$$

Layers in the residual are learning the residual  $F(x)$  and layers in the conventional are learning output  $G(x)$ . Learning both the input and the residual output rather than just the input is easier for the network. So in this way the network could learn the identity function by setting the residual to zero. Residual Network architecture hold up to 152 layers which includes pooling, convolutional and fully connected layers. So we can confidently say that this network gives accurate results when compared to other networks, provided the amount of training data and thus ResNet reduces the issue of vanishing gradients.

With the introduction of shortcut connections which are used as bypassing paths these Residual networks have achieved a striking performance on image recognition and classification tasks. Skipping has effectively increased the learning speed by reducing the impact of vanishing gradients Wikipedia contributors (2021) as there are only few layers to propagate through.

## 2.4 Dense CRF model

Depth maps which are acquired by recent CNN based depth estimation networks suffers from serious distortions at the object counters and this has still remained as a challenge in many applications. For example, if we take the depth maps that are calculated by stereo matching usually they include serious content missing and texture less areas. Depth maps acquired from physical sensors provide a robust way but however, they also suffer from some serious low resolution and noise problems which include blurs near object contours. Structure light sensors has acquired high resolution depth maps

but however,they include a large set of holes in the depth map.Some significant amount research has been done on this depth recovery based methods.We now review some of the algorithms that are based on depth recovery.

### 2.4.1 Related Work

So the RGB guided methods are classified into optimization based,learning based and filter based methods.

Optimization based methods are considered to be dominant solution for seriously distorted images.They are developed by redesigning terms based on global regularization frameworks like Conditional Random field He *et al.* (2012) and Markov Random field Ma *et al.* (2013).Zhu *et al.* (2009)proposed a robust and accurate depth estimation by considering the temporal coherence of depth maps in Markov Random field.Tao *et al.* (2017)addressed the depth in-painting and super-resolution problems based on the Conditional Random field model.Liu *et al.* (2016)proposed a robust depth map to well alleviate the boundary blurs and texture copy artifacts of recovered depth.

Learning based methods have been widely adopted for depth recovery in recent years and it provides the powerful tools for the missing contents in the seriously distorted depth images by utilizing the relevant data set.But depth structures in this methods are still not accurate and this needs to be corrected and rectified.Song *et al.* (2016) has learned the mapping from a low-resolution depth image to a high-resolution depth image by CNN, and further changed the learned depth map according to its associated RGB image. Zhang and Funkhouser (2018) finished the missing content of depth map by predicting the occlusion boundaries from RGB images.

Filter based methods are developed based on filters like guided filter He *et al.* (2012),bilateral filter Tomasi and Manduchi (1998),weighted mean as well as median filter Zhang *et al.* (2014) removed the outliers of depth map by improving the classical weighted median filter.Ma *et al.* (2013)fused both median filter and bilateral filter by well preserving the boundaries and removed the noise robustly in depth super resolution process.But however these methods do not perform well for seriously distorted maps.

In this work we recover high quality depth maps from distorted ones with structures containing errors under the guidance of RGB images.

# CHAPTER 3

## Disparity Estimator

### 3.1 Motivation

Disparity Estimation is an integral problem for Light field image processing. Most of the methods fail in images with texture less and repetitive regions. In this work we integrate deep residual network on 3D bilateral grid with dense CRF model for efficient disparity estimation.

### 3.2 Proposed network for disparity estimator

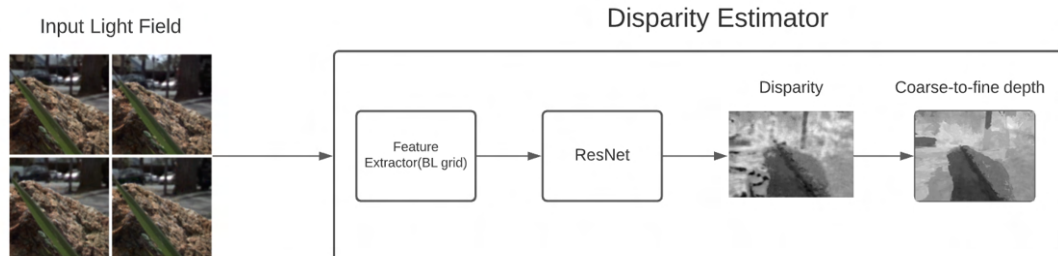


Figure 3.1: Framework of our method

The framework of our method shown in figure (3.1). The light field includes 8\*8 views and four corner views are selected as input. So the process of disparity estimator is divided into 2 steps. One is to prepare disparity features and other is to evaluate the disparity network. Out of two steps, first step is most important step and it takes a lot of time.

Here in our method we employ Bilateral grid as the feature extractor from sparse set of inputs because of its unique features like preserving and retrieving edges in the present original image and it enables fast edge aware image processing.

The architecture of our convolutional residual network is shown in figure (3.2). Our network is based on ResNet-18 layer architecture. Since a deeper ResNet layer like such

as 100-layer does not show any critical difference, compared to having little depth shallow layer in terms of the error rate, we build our architecture on ResNet-18 layer.

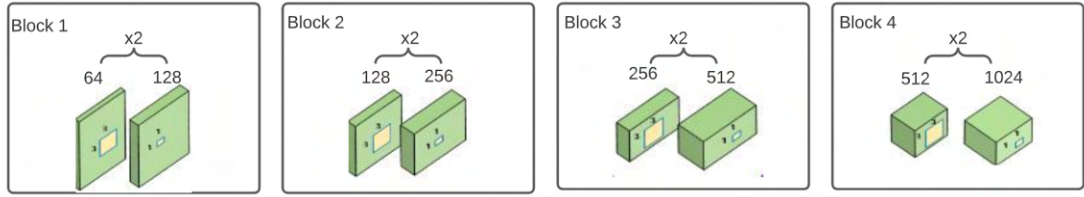


Figure 3.2: Pipeline of our convolutional Residual Network

Here we have partially added a convolutional layer at the end of each block ( $B_i \in 1, 2, 3, 4$ ) to generate a high dimensional depth map. The convolution block that is added at the end of each layer structure has same depth level with the initial layer of continuous convolution block. The depth level that has been increased will preserve the spatial information of the image which helps to generate a precise depth cost maps.

The kernel size which is used in the original convolution layer is  $3 \times 3$  and the kernel size at the end of each layer is  $1 \times 1$ . ResNet-18 in the original architecture only use  $3 \times 3$  convolutions but in our network we use an extra kernel size  $1 \times 1$  for each convolution layer at the end of each block.

We then finally implement an effective RGB guided method to recover quality depth maps from the ones with serious object contours based on dense CRF model. We address the challenging texture copy artifacts problem in the guided depth map recovery by a coarse-to-fine strategy with the pixel wise Dense CRF models. So this method can be obtained with CNN based depth estimation based or with physical sensors to obtain high quality depth maps.

### 3.2.1 Depth Structure Recovery

Depth Structure Recovery is equivalent to estimating a high quality depth map by giving a distorted map and its associated RGB image on Dense CRF model. Let  $K = K_1, \dots, K_n$  are the range of all possible depth maps to be inferred and  $K_i (i \in [1, n])$  refers to intensity of pixels  $i$  and  $n$  denotes the number of pixels in  $K$ . And similarly, let  $X = X_1, X_2, \dots, X_n$  and  $Y = Y_1, Y_2, \dots, Y_n$  denotes the distorted depth map and its associated RGB image respectively. Now the depth map can be recovered by minimizing the energy function  $E(x)$ . Now similarly,  $I$  refers to the reference image and the variable  $I_i$  refers intensity



of each pixel. The targeted image  $k = k_1, k_2, \dots, k_n$  can be inferred by maximizing the conditional probability  $k = \operatorname{argmax}_k P(\mathbf{K} = \mathbf{k} | \mathbf{I})$ .

In the Dense CRF model, this conditional probability is generally characterized by a Gibbs distribution  $P(\mathbf{K} = \mathbf{k} | \mathbf{I}) = \exp(-E(\mathbf{K} = \mathbf{k} | \mathbf{I})) / Z(\mathbf{I})$ , where  $E(\mathbf{K} = \mathbf{k} | \mathbf{I})$  is the Gibbs energy and  $Z(\mathbf{I})$  is normalization factor. So maximizing conditional probability is equal to minimizing Gibbs energy. So therefore the target image is inferred by minimizing the following energy function.

$$k = \operatorname{argmax}_k P(\mathbf{K} = \mathbf{k} | \mathbf{I}) \quad (3.1)$$

This energy function generally composed of unary and pairwise components. Where unary component intensity of pixel  $i$  in target image  $k$  independently based on reference image  $\mathbf{I}$  and pairwise component controls the spatial correlation of pixels in the target image  $k$  based on reference image  $\mathbf{I}$ .

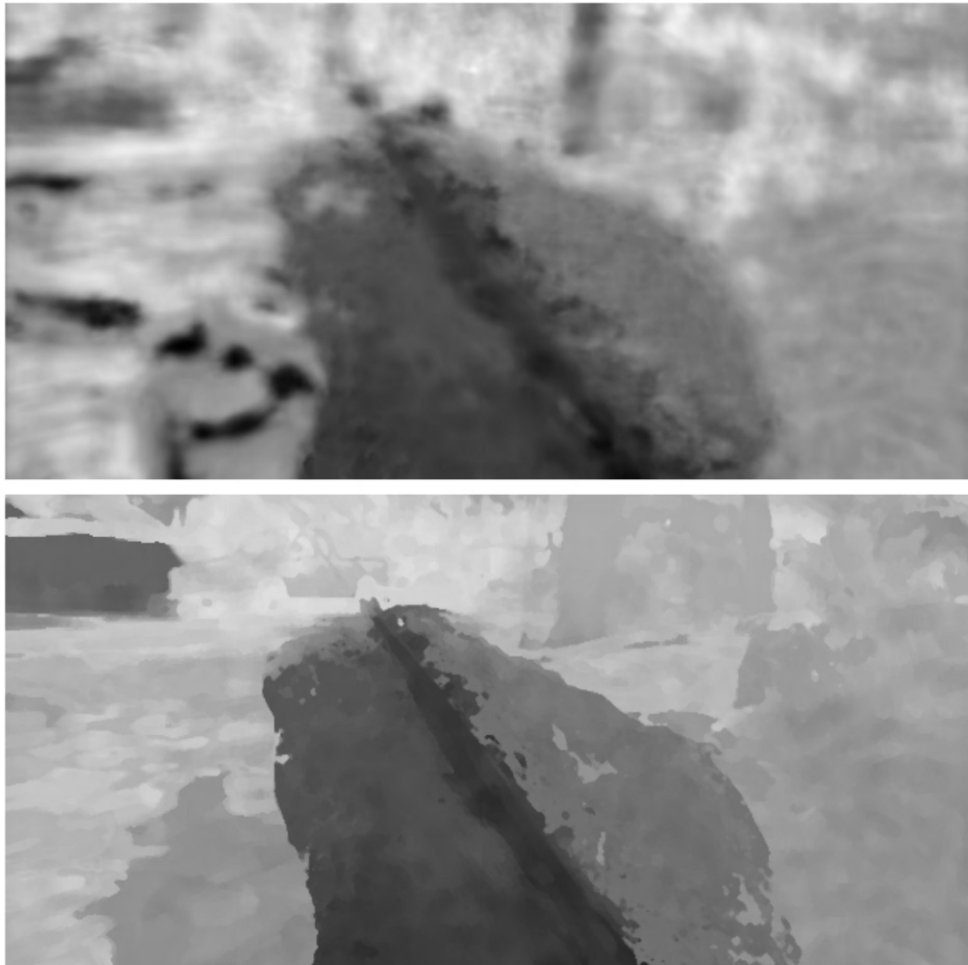


Figure 3.3: (top) distorted depth map (bottom) Coarse-to-fine depth map

The above figure(3.3) is the output of our disparity estimator network. The top image is our distorted depth map and the bottom image is our recovered quality depth map which is based on dense CRF model.

# CHAPTER 4

## Novel View Synthesis

To synthesize novel views from our estimated depth, we first map our input to the points in the point cloud. But the point cloud is however is only a partial view of the world geometry. Because of the resulting novel view can be synthesized by rendering the point cloud the novel views are incomplete with the holes caused by disocclusion. So one of the possible solution is to utilize the off-the-shelf image in-painting method to fill in the missing areas in each synthesized novel view. But this approach fails to satisfy the following requirements.

- **Temporal Consistency** - When we render multiple novel views the result needs to be temporal consistent to generate a moving camera effect. To overcome this temporal inconsistencies we independently apply an existing off the shelf in-painting method because the traditional in-painting formulations does not consider our given scenario.
- **Geometrically Consistent in-painting** - The filled-in area should resemble the background with a separation of foreground objects which occurs due to disocclusion. Existing in-painting methods do not reason about the geometry of in-painting result and which is why we are not able to satisfy this requirement.
- **Real Time Synthesis** - Best user experience is achieved when the user immediately become conscious of the results and make adjustments accordingly. Applying off-the-shield methods based on in-painting would computationally be expensive to use in this scenario.

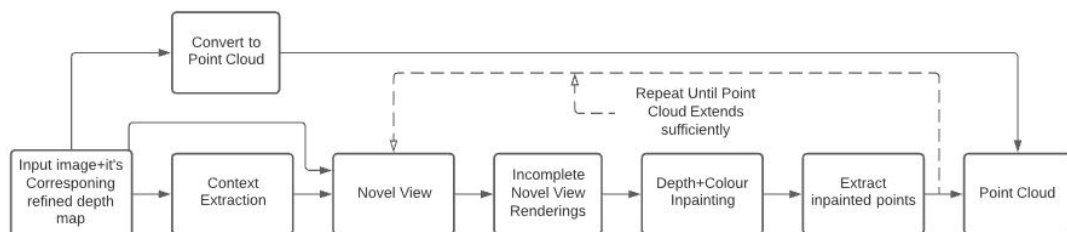


Figure 4.1: Pipeline of our Novel View Synthesis Network

Overview of our approach-The point cloud that is obtained from input image and it's corresponding refined depth map will render it's novel views from new camera

positions. Since point cloud is only a partial view of world's geometry, which is why novel views will be subject to disocclusion. So in order to address this issue we perform color and depth in-painting jointly in order to fill in the missing areas in the incomplete novel view renderings. This in-painted depth can be used to map the in-painted color in the existing point cloud by addressing the problem of occlusion. So, we repeat this procedure until the point cloud extends sufficiently to render complete and temporal consistent novel views. In this method we perform color and depth in-painting only at the extreme views like at the beginning and at the end.

Our Novel View synthesis approach is illustrated in figure ( 4.1) and we subsequently elaborate the steps involved.

#### **4.0.1 Depth and color in-painting**

Our method accepts color, depth and context information as input which is different from the existing image in-painting methods and performs depth and color in-painting jointly. The context information provides rich information that is beneficial for the high quality view synthesis. We render depth, color and context information of the input image to a novel view which is incomplete due to disocclusion. So we then use color and depth in-painting to fill in the missing areas. This in-painted depth can be used to map the in-painted color in the existing point cloud by addressing the problem of occlusion and extending the world geometry that the point cloud represents.

#### **4.0.2 Context Extraction**

Contextual Information is beneficial for generating high quality view synthesis results. Each point can be extended with the contextual information in the point cloud which describes the neighbourhood of where the corresponding pixel used to be in the input image. This helps the point cloud with rich information that can for eg., it can be leveraged for computer vision in neural rendering. So in order to make use of this technique, We leverage our network with two convolutional layers to extract 64 channels of contextual information from our input image. We train this along with the in-painting network so that it allows the extractor to allow to how to gather info that is useful when in-painting incomplete novel view renderings.

### 4.0.3 Point Cloud Rendering

Novel views can be obtained by rendering the point cloud to an image plane subject to pin hole camera model. So, while moving the virtual camera forward, the point clouds may suffer from shine through artifacts for which occluded background regions become visible in foreground regions. We identify these regions by identifying pixels for which two adjacently opposing neighbors that are significantly closer to the virtual camera.

### 4.0.4 Summary

Our novel synthesis approach addresses each of identified requirements that are required when synthesizing like Geometrical consistent in-painting, Temporal consistency and real time synthesis.

## 4.1 Comparative Analysis



Figure 4.2: Comparison of our approach against other methods

In the above figure we compare our approach against other methods using state-of-art light field disparity estimation networks as its input. We also show our estimated disparity for each scene and we can see the regions with darker color are closer to the camera. Our approach produces reasonable result which are better when compared to other methods.

## 4.2 Conclusion

In this work, we have presented a novel learning based approach for synthesizing novel views. Our system consists of depth estimator model which estimates depth from a sparse set of input views which are captured from a light field camera and a depth based view synthesis model addressing identified requirements that are required while synthesizing like temporal consistency, Geometrical consistent in-painting and real time synthesis. Experiments with other variety of image content show that our method enables realistic synthesis results. Our study shows that our methods enables users to achieve better results while taking little effort when compared to existing solutions and also shows that our approach outperforms state of art approaches.

## 4.3 Future Work

In the future, we would like to extend our network to work with any number of input views. We also would like to carry out the possibility of using our system for generating high quality light fields from a specific set of views with different kind of exposures. In the future, we would like to extend our network to work with any number of input views. We also would like to carry out the possibility of using our system for generating high quality light fields from a specific set of views with different kind of exposures. Color and depth in-painting networks that we have used in our network can be trained with real images leveraging an adverse superficial regime and a more sophisticated architecture like the one that uses partial convolutions will be an interesting work to explore in future.

## REFERENCES

1. **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.
2. **Csáji, B. C. et al.** (2001). Approximation with artificial neural networks. *Faculty of Sciences, Eötvös Loránd University, Hungary*, **24**(48), 7.
3. **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.
4. **He, K., J. Sun, and X. Tang** (2012). Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, **35**(6), 1397–1409.
5. **He, K., X. Zhang, S. Ren, and J. Sun**, Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
6. **Koch, T., L. Liebel, F. Fraundorfer, and M. Korner**, Evaluation of cnn-based single-image depth estimation methods. *In Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.
7. **LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner** (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.
8. **Liu, W., X. Chen, J. Yang, and Q. Wu** (2016). Robust color guided depth map restoration. *IEEE Transactions on Image Processing*, **26**(1), 315–327.
9. **Ma, Z., K. He, Y. Wei, J. Sun, and E. Wu**, Constant time weighted median filtering for stereo matching and beyond. *In Proceedings of the IEEE International Conference on Computer Vision*. 2013.
10. **Song, X., Y. Dai, and X. Qin**, Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network. *In Asian conference on computer vision*. Springer, 2016.
11. **Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich**, Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
12. **Tao, Q., L. Wang, D. Li, and M. Zhang** (2017). Crf-based depth refinement with hybrid depth information. *Electronics Letters*, **53**(6), 393–395.
13. **Tatarchenko, M., A. Dosovitskiy, and T. Brox** (2015). Single-view to multi-view: Reconstructing unseen views with a convolutional network. *CoRR abs/1511.06702*, **1**(2), 2.
14. **Tomasi, C. and R. Manduchi**, Bilateral filtering for gray and color images. *In Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998.

15. **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.
16. **Wikipedia contributors** (2021). Vanishing gradient problem — Wikipedia, the free encyclopedia. URL [https://en.wikipedia.org/w/index.php?title=Vanishing\\_gradient\\_problem&oldid=1006243644](https://en.wikipedia.org/w/index.php?title=Vanishing_gradient_problem&oldid=1006243644). [Online; accessed 5-June-2021].
17. **Zhang, Q., L. Xu,** and **J. Jia**, 100+ times faster weighted median filter (wmf). *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
18. **Zhang, Y.** and **T. Funkhouser**, Deep depth completion of a single rgb-d image. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
19. **Zhu, J., L. Wang, J. Gao,** and **R. Yang** (2009). Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(5), 899–909.