SUPER RESOLUTION OF LIGHT FIELDS USING CONVOLUTIONAL NEURAL NETWORK

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF

ENGINEERING AND NATURAL SCIENCES

OF ISTANBUL MEDIPOL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL, ELECTRONICS ENGINEERING AND CYBER SYSTEMS

By

Muhammad Shahzeb Khan Gul

May, 2018

ABSTRACT

SUPER RESOLUTION OF LIGHT FIELDS USING CONVOLUTIONAL NEURAL NETWORK

Muhammad Shahzeb Khan Gul
M.S. in Electrical, Electronics Engineering and Cyber Systems
Advisor: Prof. Dr. Bahadir Kursat GUNTURK
May, 2018

Light field imaging extends the traditional photography by capturing both spatial and angular distribution of light, which enables new capabilities, including post-capture refocusing, post-capture aperture control, and depth estimation from a single shot. Micro-lens array (MLA) based light field cameras offer a cost-effective approach to capture light field. A major drawback of MLA based light field cameras is low spatial resolution, which is due to the fact that a single image sensor is shared to capture both spatial and angular information. In this thesis, we present a learning based light field enhancement approach. Both spatial and angular resolution of captured light field is enhanced using convolutional neural networks. The proposed method is tested with real light field data captured with a Lytro light field camera, clearly demonstrating spatial and angular resolution improvement.

Keywords: Light field, Deep learning, Angular resolution, Spatial resolution.

ÖZET

EVRİŞİMSEL SİNİR AĞLARI İLE IŞIK ALANLARININ SÜPER ÇÖZÜNÜRLÜĞÜ

Muhammad Shahzeb Khan Gul
Elektrik-Elektronik Mühendisliği ve Siber Sistemler, Yüksek Lisans
Tez Danışmanı: Prof. Dr. Bahadır Kürşat GÜNTÜRK
Mayıs, 2018

Işık alan görüntüleme, ışığın hem uzamsal hem de açısal dağılımını kaydederek, kayıt sonrası odaklama, kayıt sonrası diyafram kontrolü ve tek bir çekimden derinlik kestirimi gibi geleneksel görüntülemeden daha öte yetenekler sağlar. Mikro-lens dizisi (MLD) tabanlı ışık alan kameraları ışık alanını kaydetmede uygun maliyetli bir yaklaşım sunar. MLD tabanlı ışık alan kameralarının temel sorunu tek bir görüntü sensörünün uzamsal ve açısal bilgiyi kaydetmesi için paylaşılmasından dolayı ortaya çıkan düşük uzamsal çözünürlüktür. Bu tezde, öğrenme temelli ışık alan iyileştirme yaklaşımı sunulmaktadır. Evrişimsel sinir ağları ile kaydedilmiş ışık alanının hem uzamsal hem de çözünürlüğü arttırılmaktadır. Önerilen metod Lytro ışık alan kamerasıyla çekilmiş gerçek ışık alan verisiyle test edilmiş, uzamsal ve açısal iyileştirme açık bir şekilde gösterilmiştir.

Anahtar sözcükler: Işık alanı, derin öğrenme, açısal çözünürlük, uzamsal çözünürlük.

Acknowledgement

First of all, I would like to thank the almighty ALLAH who gave me strength and knowledge to complete my research work. After that, I want to thank my supervisor Prof. Dr. Bahadir K. Gunturk. It would not have been possible without his constant support and knowledge. It is the result of his supervision that I am able to publish my research in one of the high impact factor journal of IEEE signal processing society.

I would also like to thank my family back in Pakistan, especially my father and mother. Their constant support and prayers gave me strength during hard times. Thank also goes to Arooba Maryam, no words in the world can express my gratitude to you for your love and sacrifice during these years. I also need to mention and thank Saba Samreen Khan, for sparing time to listen to my ideas and giving fruitful suggestions. I would also like to thank my friends Shah Rez, Tanzeel and Furqan. They helped me whenever I needed help. In the end, I would like to thank all Pakistanis here in Istanbul Medipol University for their love and support. I am extremely sorry if I forgot to mention someone by name here. The list is very long but the space is too short. Thank you, everyone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

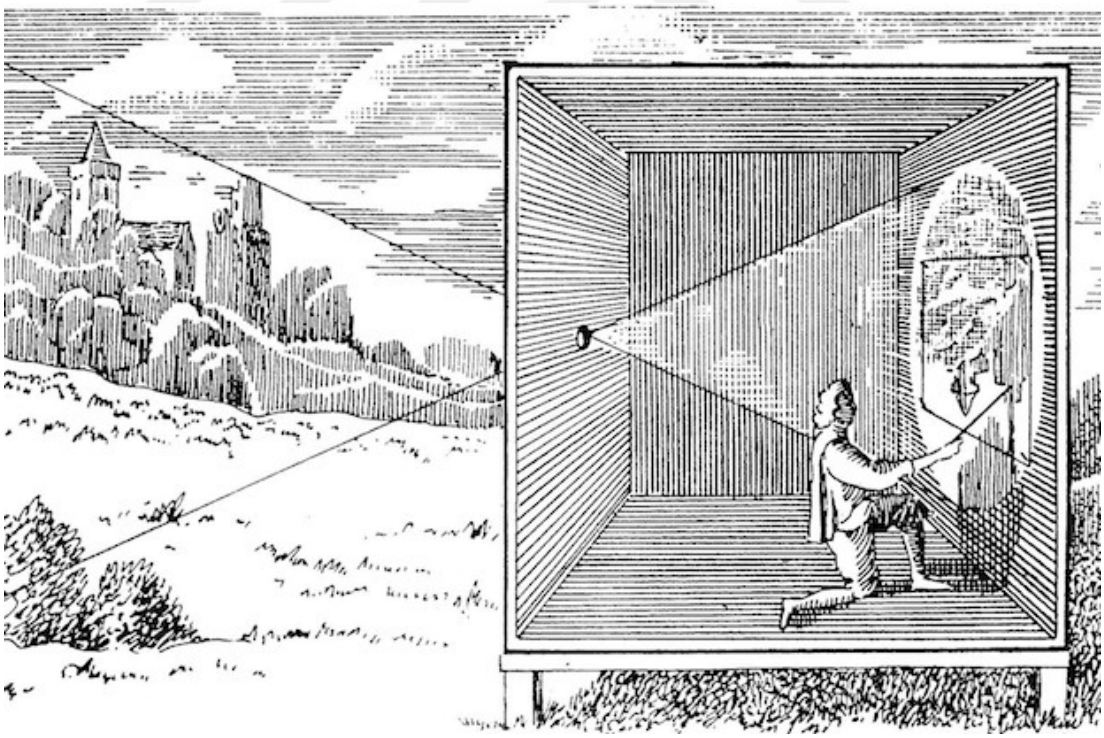


Figure 1.1: Athanasius Kircher, Large portable camera obscura, 1646. ©Gernsheim Collection.

The word "camera" is a Latin word, which means chamber or dark room. The name "camera" is derived based on the experiment performed by Ibn al-Haytham in dark room known as pinhole camera (*camera obscura*). He made a pinhole in

one wall of the dark room so that light rays can pass through it. Real world scenes reflect light rays which passes through the pinhole to form an inverted image on the opposite wall of the room. Since the inception of pinhole camera, many advancements has been made in the technology of photography.

1.1 Traditional Cameras

A traditional camera is composed of a lens system and an imager. A two-dimensional projection of a scene is formed by the lens system i.e. by converging the light rays reflected by the scene on to the imager, which records the intensities of light rays. Traditionally, photographic film or photographic plate was used as imager. The advent of digital technology introduced electronic image sensor based on charged couple device (CCD) or complementary metal-oxide-semiconductor (CMOS) technology enabling the user to save images in the digital form.

The lenses used in a lens system are usually made up of glass with refractive surfaces to converge the incident light rays. The magnitude of the refraction is dependent upon the angle between the lens surface and the light ray incident on the surface and the refractive index of the lens. To get the sharp image of the scene, the image sensor is kept at a distance where light rays are converged after passing through the lens system.

1.2 Light Field Imaging

Light field refers to the collection of light rays in 3D space. With a light field imaging system, light rays in different directions are recorded separately, unlike a traditional imaging system, where a pixel records the total amount of light received by the lens regardless of the direction. The angular light information introduce new abilities, including depth estimation, post-capture refocusing, post-capture aperture size and shape control, and 3D modeling. There are numerous application areas for light field imaging, including 3D optical inspection, robotics, microscopy, photography, and computer graphics.

Light field acquisition can be done in various ways, such as camera arrays [1], optical masks [10], angle-sensitive pixels [11], and micro-lens arrays [12, 13]. Among these diverse methodologies, micro-lens array (MLA) based light field cameras provide a cost-efficient solution, and have been successfully commercialized [14, 15].

1.3 Motivation

There are several advantages of micro-lens array (MLA) based light field cameras over the conventional camera array based light field acquisition, including compact design and cost effective solution. However, there is a fundamental trade-off between these advantages and the image resolution.

In MLA-based light field cameras, there is a trade-off between spatial resolution and angular resolution since a single image sensor is used to capture both. For example, in the first generation Lytro camera, an 11 megapixel image sensor produces an 11x11 sub-aperture perspective images, each with a spatial resolution of about 0.1 megapixels [1]. Such a low spatial resolution prevents the wide spread adoption of light field cameras. In recent years, the low spatial resolution issue has

[1]The resolution is obtained by decoding through the open-source software presented by Dansereau [16].

been widely adopted by researchers. Hybrid systems, combination of a traditional image sensor with a light field sensor, have been presented [17, 18, 19], where the high spatial resolution image from the regular sensor is used directly if needed or transfered to sub-perspective images of light field. The disadvantages of hybrid systems include increased cost and larger camera dimensions. Another approach is to apply multi-frame super-resolution techniques to the sub-perspective images of a light field [20, 21]. It is also possible to apply learning-based super-resolution techniques to each sub-aperture image of a light field [22].

1.4 Contribution

The main contribution of this thesis is to enhance the spatial and angular resolution of light fields. We have proposed a learning based method for the super-resolution of light field. The proposed method consists of the convolutional neural network for both spatial and angular resolution. The proposed method is computationally less expensive. In terms of quantitative and qualitative results, the proposed method has outperformed the state-of-the-art techniques. This work has been published as a journal paper [23].

Chapter 2

Background

2.1 Light Field

Light field imaging is first described by Lippmann, who proposed to use a set of small biconvex lenses to capture light rays in different directions and refers to it as integral imaging [24]. The term "light field" was first used by Gershun, who studied the radiometric properties of light in space [25]. Adelson and Bergen used the term "plenoptic function" and defined it as the function of light rays at every possible location in space, going at every possible angle, for every wavelength, and at every time [26]. Adelson and Wang described and implemented a light field camera that incorporates a single main lens along with a micro-lens array [27] (refer to Equation 2.2 and Figure 2.1).

$$P(\phi, \theta, x, y, z) \tag{2.1}$$

This design approach is later adopted in commercial light field cameras [14, 15]. In 1996, Levoy and Hanrahan [28] and Gortler *et al.* [29] formulated light field as a 4D function, and studied ray space representation and light field re-sampling. They restrict the attention to the light rays passing through the free-space, that is

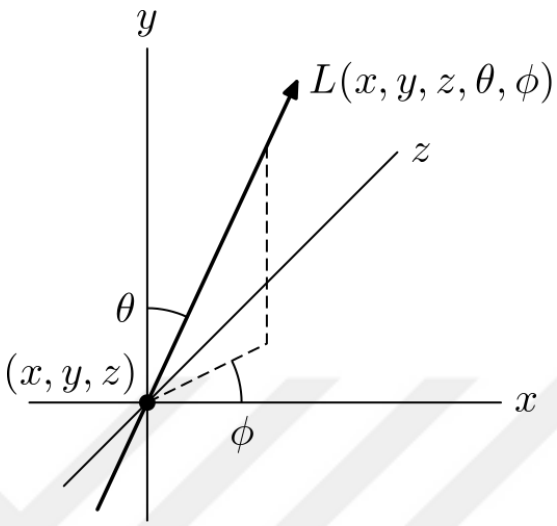


Figure 2.1: Parameterizing a ray in 3D space by position (x, y, z) and direction (θ, ϕ).

region free from occuluders. In this way, the light passing along a ray is constant and eliminating one dimension of variation.

$$P'(\phi, \theta, u, v) \tag{2.2}$$

The light intensity is given for every possible position u and v on a 2-dimensional plane, and angle θ and ϕ.

Over the years, light field imaging theory and applications have continued to be developed further. Key developments include post-capture refocusing [30], Fourier-domain light field processing [12], light field microscopy [31], focused plenoptic camera [13], and multi-focus plenoptic camera [32].

2.1.1 Light Field Parameterization

In [33], light field is represented as a 5D radiance function describing every 3D scene point through 2D light rays. However, it can be reduced to 4D with the

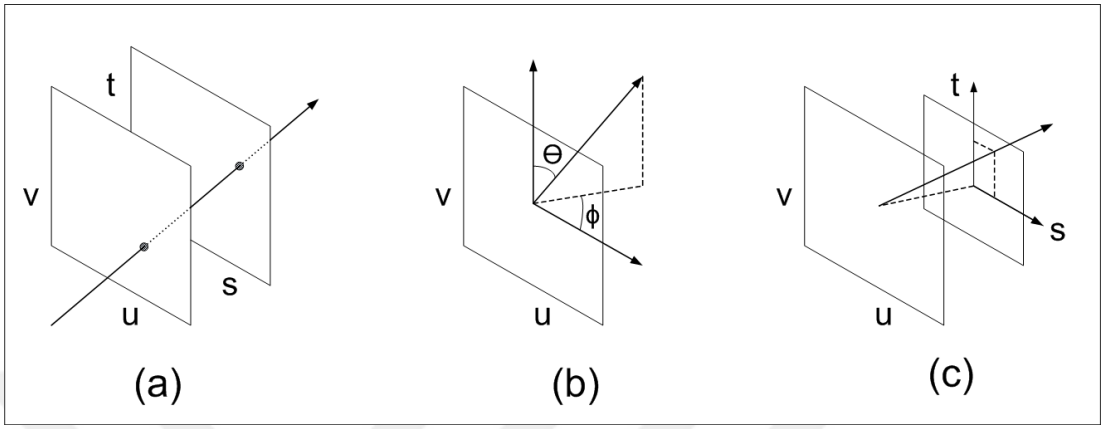


Figure 2.2: Two parallel plane light field parameterization. In all three models, u and v serves as basic arguments. The other two arguments are parameterized as; (a) global coordinates of s and t, (b) angular coordinates θ and ϕ representing the angle of ray after intersecting with uv plane, (c) local coordinates of s and t, are sometime also referred to as slope of the ray intersecting uv plane.

assumption that radiance does not change in free space along rays [28]. Parameterization of 4D light field is achieved by parameterizing light rays by their intersections with two planes in arbitrary position. In Fig. 2.2, some two-plane light field parameterization models are shown.

2.1.2 Light Field Acquisition

There are different ways to capture the light field, among them two approaches are very popular. First method is to use camera array [28, 1, 34], and the other method is to use micro-lens array in front of the imaging sensor [12, 13]. Some other light field acquisition methods include optical masks [10], angle-sensitive pixels [11], gantry for the camera movement [35], and kaleidoscope-like optics [36].

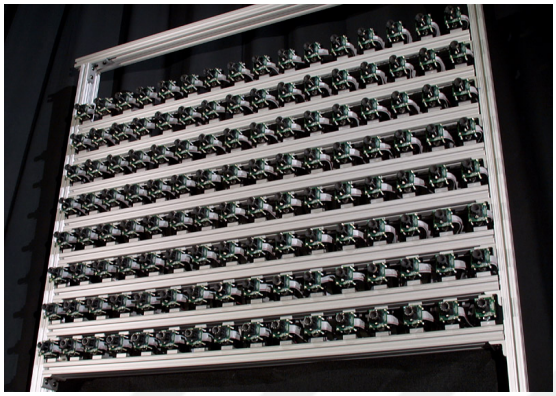


Figure 2.3: Cammera array of 8x16 cameras to create light field developed by Stanford [1].

2.1.2.1 Camera Array Based Light Field

The straightforward approach to capture high-resolution light field is to set up an array of cameras. Among different camera array-based approaches, the Stanford implementation [1] is shown in Figure 2.3. The directional information of light rays, i.e. the angular resolution is defined by the number of cameras in the grid, e.g. the angular resolution of 8x16 grid camera array as shown in Figure 2.3 is 8x16. The size of the camera array system is typically very large, such as spanning around 1m horizontally and vertically in Figure 2.3. In this case, the wider baseline between the adjacent cameras results in a discrete blur at the time of rendering of the different perspective images. The light field produced with camera array system is very high in terms of quality and high dynamic range imaging as compared to the other acquisition methods, but still, they are very bulky and not portable and sometimes required high data bandwidth.

2.1.2.2 MLA-Based Light field

MLA-based light field cameras have two basic implementation approaches. In one approach, the image sensor is placed at the focal distance of the micro-lenses [12, 14]. In the other approach, a micro-lens relays the image (formed by the objective lens on an intermediate image plane) to the image sensor [13, 15].

MLA-based light-field camera structure depicting different key components for both the approaches discussed above is shown in Fig. 2.4 and 2.5. In Fig. 2.4, the plenoptic camera 1.0 [14] design is shown, where light rays from a single scene point converges to a single point on the focal plane of the micro-lens array, from there the micro-lens separates these rays according to the direction and create a focused image of the aperture of the main lens on the grid of pixels placed exactly at the focal length of the micro-lens (which is also known as lenslet). Whereas, in the other approach plenoptic camera 2.0 [15], relative postion of the micro-lens array is the main difference. Instead of placing the micro-lenses at the principal plane of the objective lens, they are now focused onto the image plane (intermediate plane) of the objective lens. The effect of such configuration is, now each micro-lens acts as a single pinhole camera, observing the small part of the virtual image inside the camera. These small image parts are then mapped onto the image sensor with high spatial resolution. Fig. 3.1 is a dataset collected by the camera which has same structure like Fig. 2.4. Essentially the raw data is same as the conventional photograph. Microscopically, however, one can see the subimages of the objective lens aperture captured by each lenslet. These lenslet images capture the structure of light in the world, and reveal, for example, the depth of objects. The raw image formed is referred to as lenslet image.

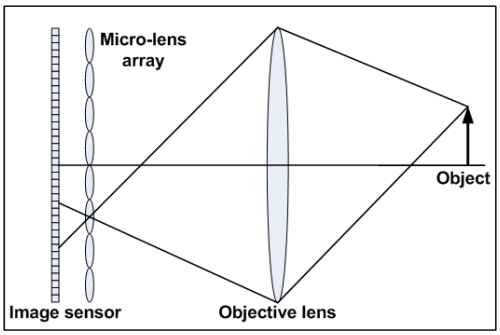


Figure 2.4: Conceptual schematic of plenoptic camera 1.0, which is composed of an objective lens, micro-lens array and image sensor. Here the image sensor is place at the focal distance of the micro-lens.

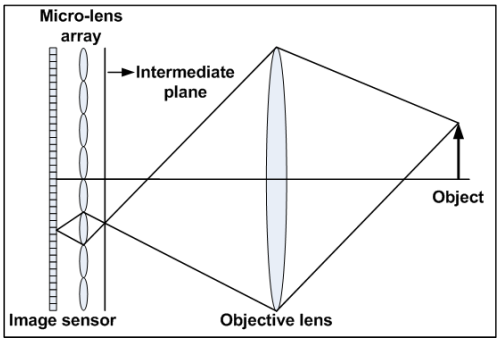


Figure 2.5: Conceptual schematic plenoptic camera 2.0, where an image formed on the intermediate plane is relays on to the image sensor.

2.2 Convolutional Neural Networks (CNNs)

In recent years, convolutional neural networks have been extensively applied in image processing and computer vision applications. CNNs are considered as a branch of deep learning algorithms. They are also known as the specific kind of neural networks designed to process data having clear grid-like topology. As the name suggests, CNNs employs convolution which is a linear mathematical operation. Although the inspiration of convolutional neural network was taken from the work of Hubel and Wiesel [37] on the identification of cells with local receptive field, but for the first time, CNN was presented in the pioneering work of Furushima's Neocognitron [38]. LeCun et al. work presented in [2] for handwriting digit recognition utilizing gradient-based stochastic gradients, is considered as the breakthrough in the field of deep learning. The architecture (for details, see Fig. 2.6) presented in [2], is the first back-propagation convolutional neural network.

CNN typically consist of several layers. These layers are further divided into different stages: convolution, with non-linear activation function and pooling operation. All these stages are explained below.

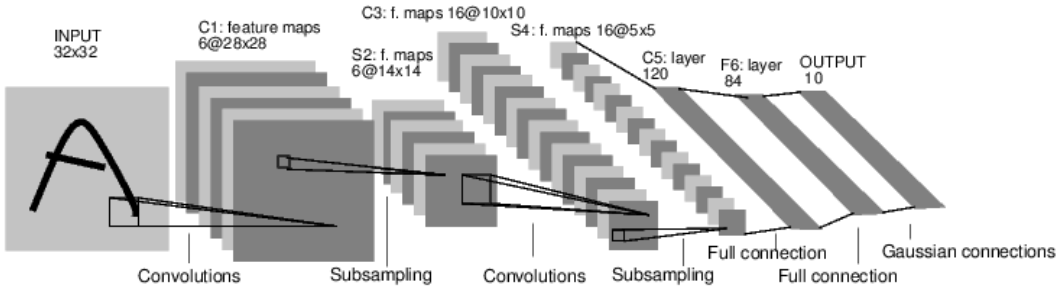


Figure 2.6: CNN model presented in [2] for document and handwritten digit recognition. This model consists of different kinds of layers such as convolution, subsampling, full connection and Gaussian connection. The input of this network is an image while the output is a vector containing the probability of the input image belonging to different classes. The image is adopted form the paper [2].

2.2.1 Convolution

The name convolutional neural network is derived due to the fact that convolution, which is a linear mathematical operation, is the key element of this algorithm. The main objective of convolution in CNN is to extract different distinct features from the input data. A two-dimensional array of learnable parameters called kernel is slide over a two-dimensional array of an input image with the summation of corresponding multiplications as output. In mathematical form, it can be expressed as

$$Y[x, y] = \sum_{i} \sum_{j} I[i, j]K[x - i, y - j] \tag{2.3}$$

Here, I denotes the input image, K is the learnable kernel and Y is the output of the convolution operation between I and K. There are three main advantages of using convolution in CNNs as compared to the matrix multiplication in multilayer neural networks. The first advantage is that it has sparse interaction (or sparse weights). It means that fewer connections between input and output which subsequently reduce the overall memory requirements. In this case, a smaller size kernel is used as compared to the size of the input. The area covered by the kernel on the input image is called receptive field. Parameter sharing is the second advantage achieved by the convolution operation. It means that there is no need to define individual kernels for every location, the same kernel is applied

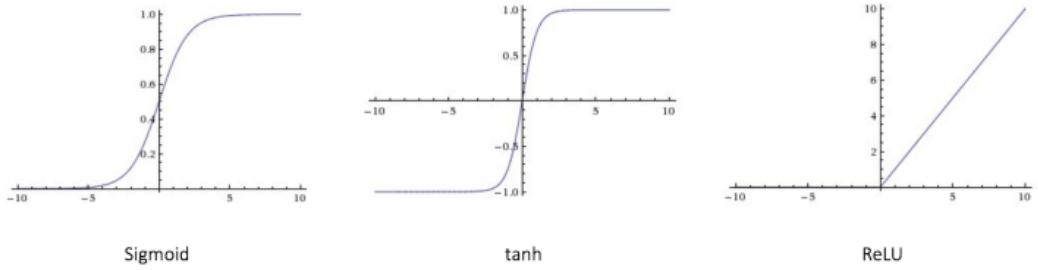


Figure 2.7: Different non-linear activation functions. The image is taken from ujjwalkarn.me

for every location in the input. Whereas, every element of the conventional neural network weight matrix is utilized only once for generating the output of the layer. This again reduces the memory requirements of the model. The last advantage is that convolution provides the equivalent representations. The output of convolution operation is a two-dimensional feature map. Parameter sharing makes the convolution operation translation equivariant. This means that the output representation changes accordingly with the variation in the input.

2.2.2 Non-linear Activation Function

Activation functions are really important to learn complex and non-linear mappings between the input and output. CNNs without non-linear activation functions would simply be a linear regression model, unable to model complicated non-linear functions. The main purpose of using CNN is to make sense of something which is complex, high dimensional and non-linear such as videos, audio, images etc. This shows the importance of activation function in CNNs. There are different kinds of non-linearities such as Sigmoid, Tanh-hyperbolic tangent and Rectified linear units (ReLU). Fig. 2.7 shows the plot of all the three functions.

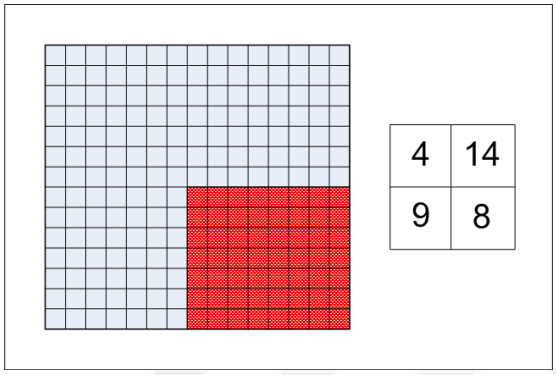


Figure 2.8: An example of the max pooling. The number of parameters are reduced by applying pooling units over four non-overlapping regions of the image.

2.2.3 Pooling

After convolution operation, a non-linear activation function, such as rectified linear activation function (ReLU) is applied on the obtained feature maps. In the last stage, pooling operation is performed. The main objective of the pooling function is to summarize and find out a single value for a 2D window. These grid windows can be overlapping and non-overlapping. The output of the max pooling layer is the maximum value in a rectangular neighborhood. Some other pooling operations are average pooling in a rectangular neighbourhood, minimum output in the neighborhood, or the $\mathcal{L}_2$ norm of the grid.

The pooling operation is again responsible for making the representation invariant to small translations of the input. This means that the small translation in the input will not affect the values of most of the pooled outputs. This invariance to translation property becomes handy when we are more interested in whether some features are present in the input rather than their exact position in the input. An example of max pooling is shown in Fig. 2.8.

An mxm region is selected to pool the convolved features obtained after convolution and non-linear activation function. Then, the convolved features are divided into mxm regions separated by s pixels. Non-overlapping regions are obtained if $s = m$. These pooled features are then utilized as input to the next convolutional layer.

2.2.4 Learning

Back-propagation algorithm or stochastic gradient descent (SGD) function is typically used in convolutional neural network to minimized a cost function. The standard formulation is given as follows

$$\theta^{k+1} = \theta^k - \frac{\eta_k \delta L(\theta^k, z)}{\delta \theta^k} \tag{2.4}$$

where θ are the learning parameters and η is the learning rate. In the recent years, computational resources have been improved due to the advancement in GPU technology. The availability of large datasets with increased computational power has let the researchers train deeper convolutional neural networks (in terms of number of layers and number of filters) to obtain significant performance increase in many computer vision applications especially for the task of image classification.

Chapter 3

Spatial And Angular Resolution Enhancement

3.1 Related Work

3.1.1 Super-Resolution of Light Field

One approach to enhance the spatial resolution of images captured with an MLA-based light field camera is to apply multi-frame super-resolution technique on the perspective images obtained from the light field capture. The Bayesian super-resolution restoration framework is commonly used, with Lambertian and textual priors [20], Gaussian mixture models [39], and variational models [21].

Learning-based single-image super-resolution methods can also be adopted to address the low spatial resolution issue of light fields. In [22], a dictionary learning based super-resolution method is presented, demonstrating a clear improvement over standard interpolation techniques when converting raw light field capture into perspective images. Another learning based method is presented in [4], which incorporates deep convolutional neural networks for spatial and angular resolution enhancement of light fields.

In contrast to single-sensor light field imaging systems, hybrid light field imaging system has also been introduced to improve spatial resolution. In the hybrid imaging system proposed by Boominathan et al. [17], a patch-based algorithm is used to super-resolve low-resolution light field views using high-resolution patches acquired from a standard high-resolution camera. There are several other hybrid imaging systems presented [18, 40, 19, 41], combining images from a standard camera and a light field camera. Among these, the work in [19, 42] demonstrates a wide baseline hybrid stereo system, increasing the spatial resolution and also improving the range and accuracy of depth estimation.

There are some methods focusing only on the low angular resolution problem associated with the light field cameras. In [43], light field super-resolution is performed from a 3D focal stack sequence using a prior based on the dimensionality gap. Frequency domain methods, utilizing signal sparsity and Fourier slice theorem is being adopted by [44], to reconstruct a 4D light field. There is another method of reconstructing full light field using multidimensional patches from a sparse set of input views [45]. An optimization framework to generate novel views from the sparse set of input views is proposed in [46]. Given the depth estimates at the input views, novel views are reconstructed by minimizing an objective function which maximizes the quality of the final results. There are some learning based methods, utilizing different convolutional neural networks architecture for the enhancement of angular resolution [3], [4]. [3] first calculate the disparity among the input views using a convolutional neural network, then utilize it to wrap the input images to the novel view using another convolutional neural network.

3.1.2 Deep Learning for Image Restoration

Convolutional neural networks (CNNs) are variants of multi-layer perceptron networks. Convolution layer, which is inspired from the work of Hubel and Wiesel [37] showing that visual neurons respond to local regions, is the fundamental part of a CNN. In [2], LeCun et al. presented a convolutional neural network based

Figure 3.1: Light field captured by a Lytro Illum camera. A zoomed-in region is overlaid to show the individual lenslet regions.

pattern recognition algorithm, promoting further research in this field. Deep learning with convolutional neural networks has been extensively and successfully applied to computer vision applications. While most of these applications are on classification and object recognition, there are also deep-learning based low-level vision applications, including compression artifact reduction [47], image deblurring [48] [49], image deconvolution [50], image denoising [51], image inpainting [52], removing dirt/rain noise [53], edge-aware filters [54], image colorization [55], and in medical imaging for the automatic segmentation of retinal layer boundaries in optical coherence tomography (OCT) images [56], which is the first step in creating high quality vascular pattern images from the popular new OCT angiography imaging modalities [57, 58]. Recently, CNN has also been deployed for the super-resolution of single image [59, 60, 61, 62]. Although these single-frame super-resolution methods can be directly applied to light field perspective images to improve their spatial resolution, we expect better performance if the angular information available in the light field data is also exploited.

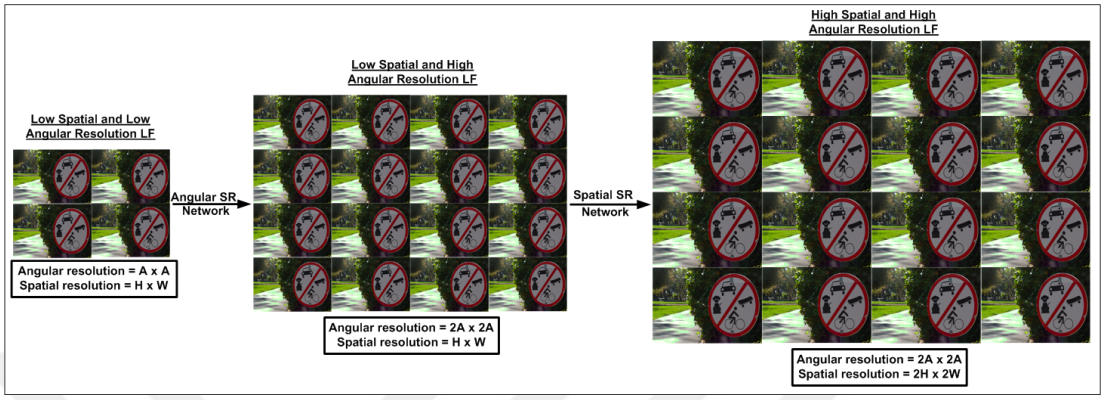


Figure 3.2: An illustration of the proposed DLFSR method. First, the angular resolution is doubled; second, the spatial resolution is doubled. The networks are applied directly on the raw light field, not on the perspective images. The effect of each step on the perspective images is also illustrated.

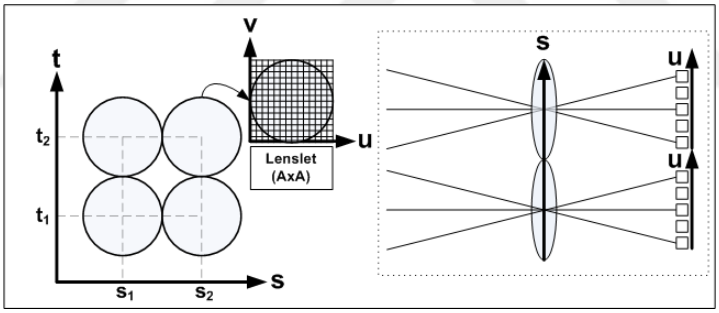


Figure 3.3: Light field parameterization. Light field can be parameterized by the lenslet positions *(s,t)* and the pixel positions *(u,v)* behind a lenslet.

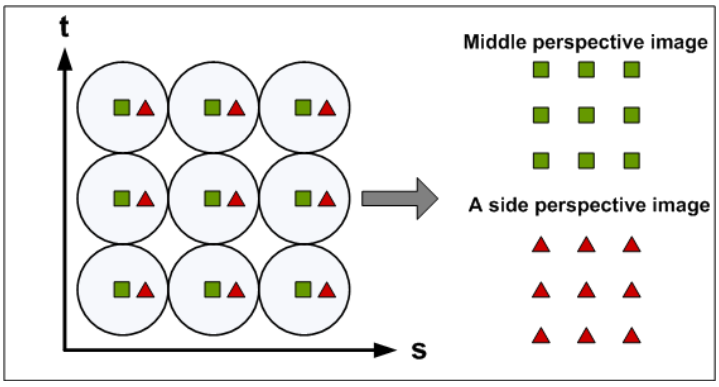


Figure 3.4: Sub-aperture (perspective) image formation. A perspective image can be constructed by picking specific pixels from the lenslet regions. The size of a perspective image is determined by the number of lenslets.

3.2 Architecture and Formulation

In Figure 3.1, a light field captured by a micro-lens array based light field camera (Lytro Illum) is shown. When zoomed-in, individual lenslet regions of the MLA can be seen. The pixels behind a lenslet region record directional light amounts received by that lenslet. As illustrated in Figure 3.3, it is possible to represent a light field with four parameters *(s,t,u,v)*, where *(s,t)* indicates the lenslet location, and *(u,v)* indicates the angular position behind the lenslet. A perspective image can be constructed by taking a single pixel value with a specific *(u,v)* index from each lenslet. The process is illustrated in Figure 3.4. The spatial resolution of a perspective image is controlled by the size and the number of the lenslets. Given a fixed image sensor size, the spatial resolution can be increased by having smaller size lenslets; given a fixed lenslet size, the spatial resolution can be increased by increasing the number of lenslets, thus, the size of the image sensor. The angular resolution, on the other hand, is defined by the number of pixels behind a lenslet region.

Our goal is to increase both spatial and angular resolution of a light field capture. We propose a convolutional neural network based learning method, which we call *light field super resolution* (LFSR). It consists of two steps. Given a light field where there are AxA pixels in each lenslet area and the size of each perspective is HxW, the first step doubles the angular resolution from AxA to $2A$x$2A$ using a convolutional neural network. In the second step, the spatial resolution is doubled from HxW to $2H$x$2W$ by estimating new lenslet regions between given lenslet regions. Figure 3.2 gives an illustration of these steps.

The closest work in the literature to our method is the one presented in [4], which also uses deep convolutional networks. There is a fundamental difference between our approach and the one in [4]; while our architecture is designed to work on raw light field data, that is, lenslet images; [4] is designed to work on perspective images. In the experimental results section, we provide both visual and quantitative comparison with [4].

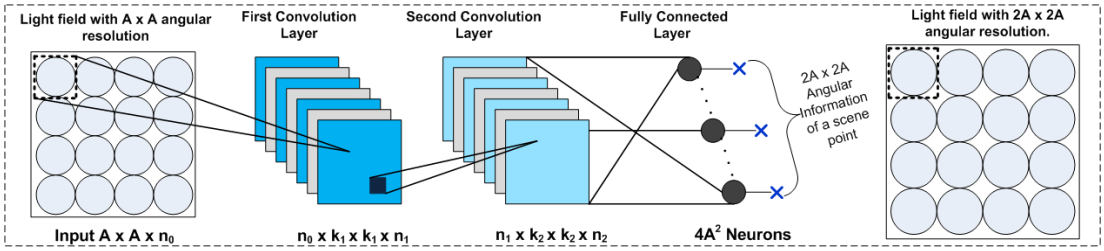


Figure 3.5: Overview of the angular SR network to estimate a high angular resolution version of the input light field. A lenslet is drawn as a circle; the AxA region behind a lenslet is taken as the input and processed to predict a $2A$x$2A$ lenslet region. ReLU non-linear activation function is applied after each convolution layer.

3.2.1 Angular Super-Resolution (SR) Network

The proposed angular super-resolution network is shown in Figure 3.5. It is composed of two convolutional layers and a fully connected layer. The input to the network is a lenslet region with size AxA; and the output is a higher resolution lenslet region with size $2A$x$2A$. That is, the angular resolution enhancement is done directly on the raw light field (after demosaicking) as opposed to doing on perspective images. Each lenslet region is interpolated by applying the same network. Once the lenslet regions are interpolated, one can construct the perspective images by rearranging the pixels, as mentioned before. At the end, $2A$x$2A$ perspective images are obtained from AxA perspective images.

The convolution layers in the proposed architecture are based on the intuition that the first layer extracts a high-dimensional feature vector from the lenslet and the second convolutional layer maps it onto another high-dimensional vector. After each convolution layer, there is a non-linear activation layer of *Rectified Linear Unit* (ReLU). In the end, a fully connected layer aggregates the information of the last convolution layer and predicts a high-resolution version of the lenslet region.

The first convolution layer has n_1 filters, each with size n_0xk_1xk_1. (In our experiments, we treat each color channel separately, thus $n_0 = 1$.) The second convolution layer has n_2 filters, each with size n_1xk_2xk_2. The final layer is a fully

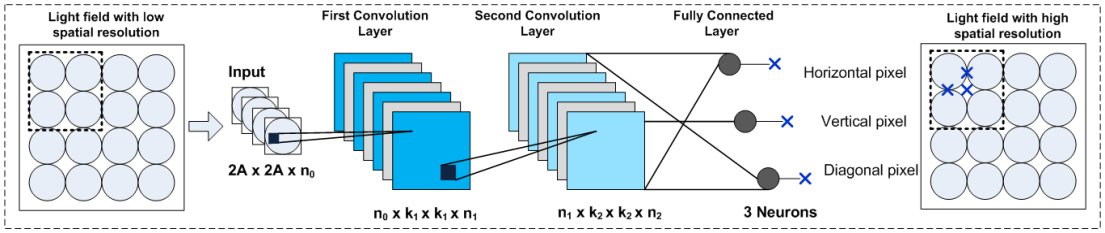


Figure 3.6: Overview of the proposed spatial SR network to estimate a higher spatial resolution version of the input light field. Four lenslet regions are stacked and taken as the input to the network. The network predicts three new pixels to be used in high-resolution perspective image formation. Each convolution layer is followed by a non-linear activation layer of ReLU.

connected layer with $4A^2$ neurons, forming a $2A$x$2A$ lenslet region.

3.2.2 Spatial Super-Resolution (SR) Network

Figure 3.6 gives an illustration of the spatial super-resolution network. Similar to the angular super-resolution network, the architecture has two convolutional layers, each followed by a ReLU layer, followed by a fully connected layer. Different from the angular resolution network, four lenslet regions are stacked together as the input to the network. There are three outputs at the end, predicting the horizontal, vertical, and diagonal sub-pixels of a perspective image. To clarify the idea further, Figure 3.7 illustrates the formation of a high-resolution perspective image. As mentioned earlier, a perspective image of a light field is formed by picking a specific pixel from each lenslet region and putting all picked pixels together according to their perspective lenslet positions. Using four lenslet regions, the network predicts three additional pixels in between the pixels picked from the lenslet regions. The predicted pixels, along with the picked pixels, form a higher resolution perspective image.

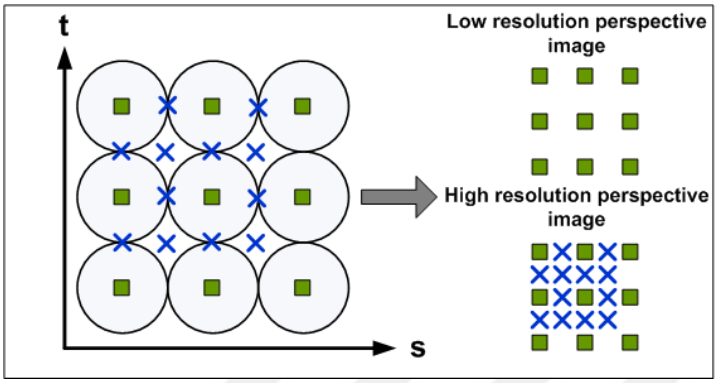


Figure 3.7: Constructing a high-resolution perspective image. A perspective image can be formed by picking a specific pixel from each lenslet region, and putting all picked pixels together. Using the additional pixels predicted by the spatial SR network, a higher resolution perspective image is formed.

3.3 Training

We used a dataset that is captured by a Lytro Illum camera and is available online [63]. The dataset has more than 200 raw light fields, each with angular resolution of 14x14 and spatial resolution of 374x540. In other words, each light field consists of 14x14 perspective images; and each perspective image has a spatial resolution of 374x540 pixels. The raw light field is of size 5236x7560, consisting of 374x540 lenslet regions, where each lenslet region has 14x14 pixels. We used 45 light fields for training and reserved the others for testing. The training data is obtained in two steps. First, we drop every other lenslet region to obtain a low-spatial-resolution (187x270) and high-angular-resolution (14x14) light field. Second, we drop every other pixel in a lenslet region to obtain a low-spatial-resolution (187x270) and low-angular-resolution (7x7) light field.

The angular SR network, as shown in Figure 3.5, has low-spatial-resolution and low-angular-resolution light field as its input, and low-spatial-resolution and high-angular-resolution light field as its output. Each lenslet region is treated separately by the network, increasing the size from 7x7 to 14x14. The first convolution layer consists of 64 filters, each with size 1x3x3. It is followed by a ReLU layer. The second convolution layer consists of 32 filters of size 64x1x1, followed by a ReLU layer. Finally, there is a fully connected layer with 196 neurons at the

end to produce a 14x14 lenslet region.

The spatial super-resolution network, as shown in Figure 3.6, has low-spatial-resolution and high-angular-resolution light field as its input, and high-spatial-resolution and high-angular-resolution light field as its output. Four lenslet regions are stacked to form a 14x14x4 input. The first convolution layer consists of 64 filters, each with size 4x3x3. The second convolution layer consists of 32 filters of size 64x1x1. Each convolution layer is followed by a ReLU layer. Finally, there is a fully connected layer with three neurons at the end to produce the horizontal, vertical and diagonal pixels. This network generates one high-spatial resolution perspective. For each perspective, the network is trained separately.

We implement and train our model using the Caffe package [64]. For the weight initialization of both networks, we used Xavier's initialization technique [65], with mean set to zero and standard deviation set to 10^{-3}, to prevent vanishment or over-amplification of weights. The learning rate for the three layers of the networks is 10^{-3}, 10^{-3}, and 10^{-5}, respectively. Mean squared error is used as the loss function, which is minimized using the stochastic gradient descent method with standard backpropagation [2]. For each network, the input size is about 13 million; and the number of iterations is about 10^8.

Chapter 4

Experiments

We evaluated our LFSR method on 25 test light fields which we reserved from the Lytro Illum camera dataset [63] and on the HCI dataset [66]. For spatial and angular resolution enhancement, we compared our method against the LFCNN [4] method and bicubic interpolation. There are several methods in the literature that synthesize new viewpoints from a light field data; thus, we compared the angular SR network of our method with two such view synthesis methods, namely, Kalantari et al. [3] and Wanner and Goldluecke [46]. Finally, there are single-frame spatial resolution enhancement methods; we chose the latest state-of-the-art method, called DRRN [67], and included it in our comparisons.

In addition to spatial and angular resolution enhancement, we investigated depth estimation performance, and compared the depth maps generated by low-resolution light fields and the resolution-enhanced light fields. In the end, we investigated the effect of the network parameters, including the filter size and the number of layers, on the performance of the proposed spatial SR network.

4.1 Dataset

4.2 Spatial Resolution

The test images are downsampled from 1414 perspective images, each with size 374x 540 pixels, to 7x7 perspective images with size 187x270 pixels by dropping every other lenslet region and every pixel in each lenslet region. The trained networks are applied to these low-spatial and low-angular resolution images to bring them back to the original spatial and angular resolutions. The networks are applied on each color channel separately. Since the original perspective images available, we can quantitatively calculate the performance by comparing the estimated and the original images. In Table 4.1, we provide peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [68] results of our method, in addition to the results of the LFCNN [4] method and bicubic interpolation. Here, we should make two notes about the LFCNN method. First, we took the learned parameters provided in the original paper and fine tuned them with our dataset as described in [4]. This revision improves the performance of the LFCNN method for our dataset. Second, the LFCNN method is designed to split a low-resolution image pixel into four sub-pixels to produce a high-resolution image; therefore, we included the results of bicubic resizing (imresize function in MATLAB) to evaluate the quantitative performance of the LFCNN method. In Table 4.1, we see that the LFCNN method produces about 1.3 dB better than the bicubic resizing. The proposed method produces the best results in terms of PSNR and SSIM.

Visual comparison is critical when evaluating spatial resolution enhancement. Figures 4.6, 4.7, and 4.9 are typical results from the test dataset. Figure 12 is our worst result among all test images. In these figures, we also include the results of the single-image spatial resolution method, called DRRN [67]. This method is based on deep recursive residual network technique, and produces state-of-the-art results in spatial resolution enhancement. Examining the results visually, we

Methods	PSNR (dB)			SSIM		
	Min	Avg	Max	Min	Avg	Max
Bicubic resizing(*imresize*)	24.2029	27.6671	34.6330	0.7869	0.8744	0.9457
LFCNN [4]	25.5963	28.9661	34.8231	0.7838	0.8904	0.9407
Bicubic interp.	27.2620	30.6245	37.1640	0.5780	0.9256	0.9659
Proposed (LFSR)	**29.7515**	**33.4273**	**39.5655**	**0.9360**	**0.9559**	**0.9823**

Table 4.1: Comparison of different spatial and angular resolution enhancement methods.

conclude that our method performs better than LFCNN method and bicubic interpolation, and produces comparable results with the DRNN method. We notice that the LFCNN method produces sharper results compared to bicubic interpolation despite having lower PSNR values. In our worst result, given in Figure 4.11, the DRNN method outperforms all methods. This particular image has highly complex texture, which seems to be not modeled well with the proposed architecture. Training with similar images or using more complex architecture may improve the performance. When comparing deep networks, we should consider the computational cost as well. The computation time for one image with the DRRN method is about 859 seconds, whereas, the proposed SR network takes about 53 seconds, noting that both are implemented in MATLAB on the same machine

In Figure 4.10, we test our method on the HCI dataset [66]. We compare against the networks in [4] and [5]. The method in [5] produces less ringing artifacts compared to the LFCNN network [4]. The proposed method again produces the best visual results.

Although we have shown results for resolution enhancement of the middle perspective image so far, the proposed spatial SR network can be used for any perspective image as well. In Table 4.2, average PSNR and SSIM on test images for different perspective images (among the 14x14 set) are presented. It is seen that similar results are obtained on all perspective images, as expected.

Sub-aperture Image#	Method	PSNR (dB)	SSIM
7,1	Bicubic interp.	28.21	0.8631
	Proposed	**28.74**	**0.8789**
5,5	Bicubic interp.	31.12	0.9310
	Proposed	**32.95**	**0.9496**
6,6	Bicubic interp.	30.74	0.9272
	Proposed	**32.71**	**0.9485**
6x8	Bicubic interp.	30.73	0.9267
	Proposed	**32.94**	**0.9504**
8,6	Bicubic interp.	30.69	0.9270
	Proposed	**32.69**	**0.9492**
8,8	Bicubic interp.	27.71	0.8793
	Proposed	**28.16**	**0.8917**

Table 4.2: Evaluation of the proposed method for different perspective images.

4.3 Angular Resolution

In this section, we evaluate the individual performance of our angular SR network. For this experiment, the angular resolution of the test images are downsampled from 14x14 to 7x7 while keeping the spatial resolution at 374x540 pixels. These low-angular images are then input to the angular SR network to bring them back to the original angular resolution. The network is trained for each color channel separately. We compare our method against Kalantari et al. [3], which is a very recent convolutional neural network based novel view synthesis method, and against Wanner and Goldluecke [46], which utilizes disparity maps in a variational optimization framework. Wanner and Goldluecke [46] may work with any disparity map generation algorithm; thus, we report results with the disparity generation algorithms given in [6], [7], [8], and [9]. In Table 4.7, we quantitatively compared the results with the state-of-the-art angular resolution enhancement methods using PSNR and SSIM. In Figure 4.12, we provide a visual comparison. Occluded regions in the scene increases the difficulty for view synthesis. The proposed angular SR method produces significantly better results compared to all other approaches.

Finally, we would like to note that the angular SR network, by itself, may

turn out to be useful, since it may be combined with any single-image resolution enhancement method to enhance the spatial and angular resolution of a light field capture.

4.3.1 Lenslets Grid VS Single Lenslet Input

The input of our Angular SR network is a single lenslet. In this section, we have carried out an experiment by increasing the number of lenslet form one to a grid of 3x3 i.e. total nine lenslets. As shown in Fig. 4.1, the overall architecture formation of the new network is exactly same as the angular SR network. To distinguish between the two network, the angular SR network taking a grid of lenslets as input is called *Grid Angular SR network*.

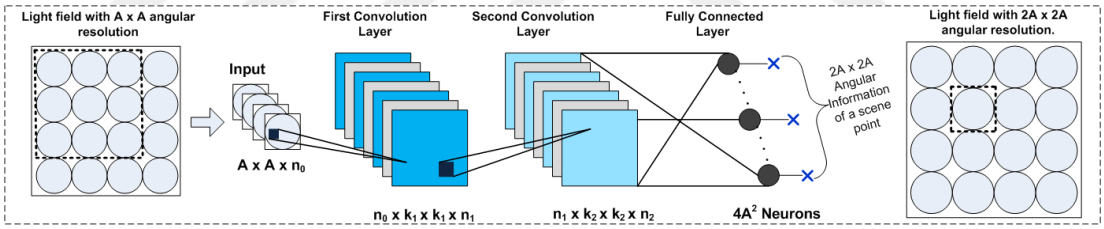


Figure 4.1: Architecture of the proposed grid angular SR network to estimate a higher angular resolution version of the input light field using lenslets grid as input. Nine lenslet regions are stacked (i.e. AxAx9) and taken as the input to the network. The output of the network is $2A$x$2A$ middle lenslet region. ReLU non-activation function is applied after each convolution layer.

For the training of the grid angular SR network, we used the same dataset, captured using a Lytro Illum camera [63]. The network takes a grid of 3x3 lenslet regions stacked to form 7x7x9 input, and increase the size of the middle input lenslet region from 7x7 to 14x14.

In Table 4.4 and 4.3, we quantitatively compared the results with the Kalantari *et al.* [3] angular resolution enhancement technique using PSNR (peak-signal-to-noise-ration) and SSIM (structural similarity). The results are presented for the two test datasets, the one used in this thesis and the other one is presented in [3], which contains challenging diverse light fields. Both the proposed angular SR

networks have shown significant improvement as compared to the state-of-the-art method achieving high PSNR and SSIM. Although the overall difference in PSNR and SSIM values between both angular SR network and grid angular SR network is not very high. The computational time for a light field for angular SR network is 66.01 sec, whereas the time taken by grid angular SR network is 73.83 sec.

Method	PSNR (dB)	SSIM
Kalantari *et al.* [3]	37.50	0.970
Angular SR Network	41.96	0.9846
Grid Angular SR Network	**42.68**	**0.9884**

Table 4.3: Quantitative evaluation of different angular resolution enhancement methods on test dataset provided in [3] containing 30 light fields.

Method	PSNR (dB)	SSIM
Kalantari *et al.* [3]	32.33	0.9339
Angular SR Network	41.25	0.9904
Grid Angular SR Network	**42.15**	**0.9912**

Table 4.4: Quantitative evaluation of different angular resolution enhancement methods on test dataset used in this thesis, containing 25 light fields.

In Figure 4.2, we have provided visual comparison. Both the images are very challenging due to the occlusion and complex structure. It is clearly visible that both proposed networks have produced visually same results, and better than the state-of-the-art method.

4.4 Depth Estimation

One of the capabilities of light field imaging is depth map estimation, whose accuracy is directly related to the angular resolution of light field. In Figure 4.8, and 4.3, we compare depth maps obtained from the input light fields and the light fields enhanced by the proposed method. The depth maps are estimated using the method in [9], which is specifically designed for light fields. It is clearly visible that depth maps obtained from light fields enhanced with the proposed method show higher accuracy. With the enhanced light fields, even close depths can be differentiated, unlike the low-resolution light fields.

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.2: Visual comparison of the novel view. Both the pictures ("Cars" and "Flower2") are taken from Kalantari *et al.* [3] paper. (a) Kalantari *et al.* [3]/31.65 dB. (b) Angular SR Network/35.21 dB (c) Grid Angular SR Network/35.42 dB (d) Ground truth, of "Cars" picture. (e) Kalantari *et al.* [3]/31.93 dB. (f) Angular SR Network/36.75 dB (g) Grid Angular SR Network/37.42 dB (h) Ground truth, of "Flower2" picture.

4.5 Models and Performance Tradeoff

To evaluate the best trade off between performance and speed, and to investigate the relations between performance and the network parameters, we will progressively modify different parameters from proposed network settings (i.e. $k_1 = 3$, $k_2 = 1$, $n_0 = 1$, $n_1 = 64$, $n_2 = 32$ and $n_3 = 3$). All the experiments are performed on a machine with Intel Xeon CPU E5-1650 v3 3.5GHz, 16GB RAM and Nvidia 980ti 6GB graphics card.

(a) (b) (c)

Figure 4.3: Depth map estimation accuracy. (a) Middle perspective image. (b) Estimated depth from the input light field of 7x7 angular resolution. (c) Estimated depth from enhanced light field with 14x14 angular resolution.

4.5.1 Filter Size

In the proposed spatial SR network, the filter sizes in the two convolution layers are $k_1 = 3$ and $k_2 = 1$, respectively. The filter size of the first convolution layer is kept at $k_1 = 3$; this means, for each light ray (equivalently, perspective image), the network is considering the light rays (perspective images) in a 3x3 neighborhood in the first convolution layer. Since higher dimensional relations are taken care of in the second convolution layer, and since keeping the filter size small minimizes the boundary effects, note that the input size in the first layer is 14x14, this seems to be a reasonable choice for the first layer. On the other hand, we have more flexibility in the second convolution layer. We examined the effect of the filter size in the second convolution layer by setting $k_2 = 3$ and $k_2 = 5$ while keeping the other parameters intact. In Figure 4.4, we provide the average PSNR values on the test dataset for different values of k_2 as a function of training backpropagation numbers. When $k_2 = 5$, the convergence is slightly better than the case with $k_2 = 1$. In Table 4.5, we show the final PSNR values and the computation times per channel (namely, the red channel) for a perspective image. It is seen that while the PSNR is slightly improved the computation time is more than doubled when we increase the filter size from $k_2 = 1$ to $k_2 = 5$.

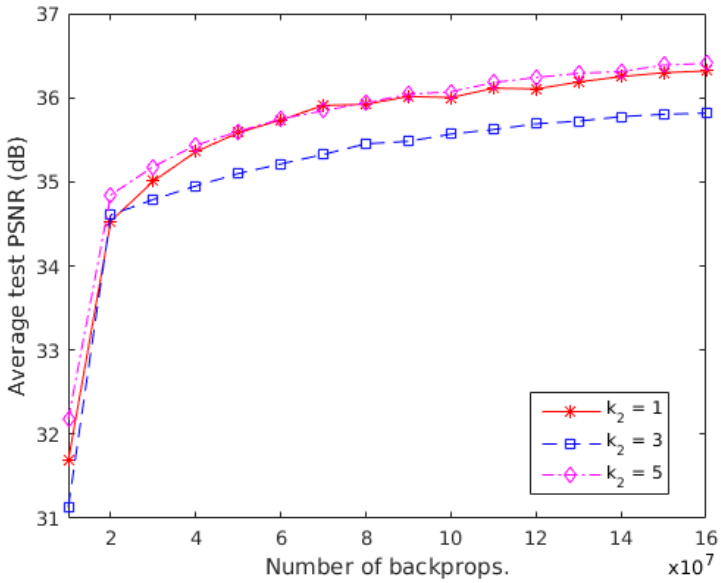


Figure 4.4: Effect of the filter size on performance.

Filter size	PSNR(dB)	Time(sec)
$k_2 = 1$	36.31	17.58
$k_2 = 3$	35.81	27.62
$k_2 = 5$	36.40	45.18

Table 4.5: Effect of the filter size on performance and speed of the spatial SR network.

4.5.2 Number of Layers and Numer of Filters

We also examine the network performance for different number of layers and different number of filters. We implemented deeper architectures by adding new convolution layers after the second convolution layer. The three-layer network presented in the previous section is compared against the four-layer and five-layer networks. For the four-layer network, we evaluated the performance for different filter combinations. The network configurations we used are shown in Table 4.6. In Figure 4.5, we provide the convergence curves for these different network configurations. We observe that the simple three-layer network performs better than the others. This means that increasing the number of convolution layers is causing overfitting and degrading the performance.

	Convolutional Layers			
	First	Second	Third	Forth
3 layer	1x3x3x64	64x1x1x32	-	-
4 layer	1x3x3x64	64x1x1x32	32x1x1x32	-
4 layer	1x3x3x64	64x1x1x16	16x1x1x16	-
4 layer	1x3x3x64	64x1x1x32	32x1x1x16	-
5 layer	1x3x3x64	64x1x1x16	16x1x1x16	16x1x1x16

Table 4.6: Different network configurations used to evaluate the performance of the spatial SR network.

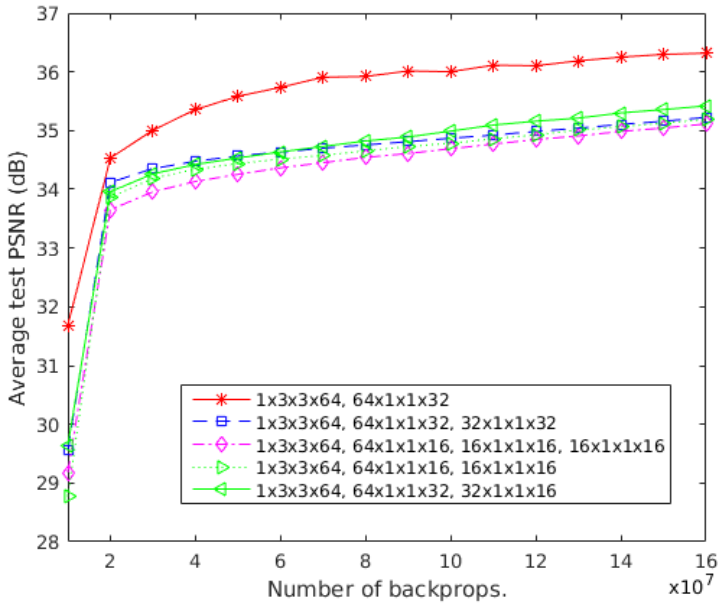


Figure 4.5: Effect of number of layers and number of filters on performance.

(a) Ground Truth | (b) Bicubic resizing(*imresize*) / 25.34 dB | (c) Bicubic interp. / 27.7 dB | (d) LFCNN [4]/ 25.76 dB | (e) DRRN [67]/ 31.63 dB | (f) LFSR / 32.25 dB

Figure 4.6: Visual comparison of middle sub-aperture image.

(a) Ground Truth | (b) Bicubic resizing(*imresize*) / 25 dB | (c) Bicubic interp. / 28.11 dB | (d) LFCNN [4]/ 25.12 dB | (e) DRRN [67]/ 32.20 dB | (f) LFSR / 32.35 dB

Figure 4.7: Visual comparison of middle sub-aperture image.

(a) (b) (c)

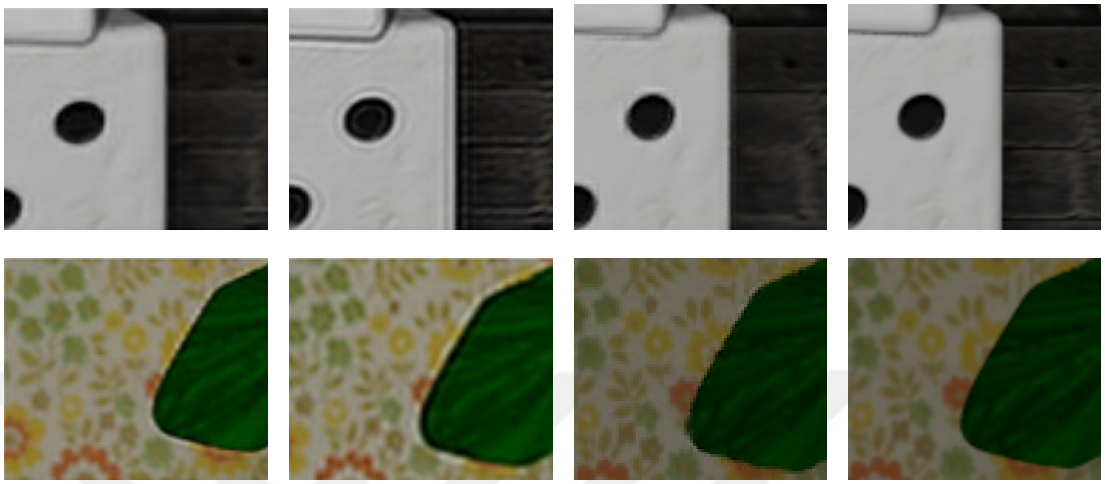
Figure 4.8: Depth map estimation accuracy. (a) Middle perspective image. (b) Estimated depth from the input light field with 7x7 angular resolution. (c) Estimated depth from enhanced light field of 14x14 angular resolution.

(a) Ground Truth | (b) Bicubic resizing(*imresize*) / 28.92 dB | (c) Bicubic interp. / 28.11 dB | (d) LFCNN [4]/ 29.06 dB | (e) DRRN [67]/ 36.78 dB | (f) LFSR / 34.21 dB

Figure 4.9: Visual comparison of middle sub-aperture image.

Name	Evaluation metrics	Wanner *et al.* [46] wanner *et al.* [6]	Tao *et al.* [7]	Wang *et al.* [8]	Jeon *et al.* [9]	Kalantari *et al.* [3]	Angular SR network
FLOWER 1	PSNR(dB)	22.03	29.52	24.39	28.21	33.31	**35.95**
	SSIM	0.789	0.941	0.910	0.934	0.969	**0.982**
CARS	PSNR(dB)	19.74	27.27	22.09	27.51	31.65	**35.21**
	SSIM	0.792	0.946	0.911	0.949	0.966	**0.983**
FLOWER 2	PSNR(dB)	20.61	27.56	23.65	27.04	31.93	**36.75**
	SSIM	0.645	0.919	0.899	0.924	0.959	**0.980**
ROCK	PSNR(dB)	16.57	30.46	30.55	30.21	**34.67**	34.09
	SSIM	0.488	0.945	0.948	0.946	**0.970**	0.963
LEAVES	PSNR(dB)	15.03	23.54	20.08	23.88	27.80	**33.08**
	SSIM	0.481	0.882	0.855	0.893	**0.963**	0.956

Table 4.7: Quantitaive comparison of different methods for angular resolution enhancement.



(a) (b) (c) (d)

Figure 4.10: Qualitative comparison of generated novel views of HCI dataset. Compared with LFCNN network presented in [4] showing ringing artifacts in high-frequency regions, while the result of the LFCNN network presented in [5] has much less artifacts. On the other hand the proposed method produce result which are very close to the ground truth. (a) Yoon et al. [5]. (b) Yoon et al. [4]. (c) Proposed. (d) Ground truth.

4.6 Further Increasing the Spatial Resolution

For quantitative evaluation, we need to have the ground truth; thus, we downsample the captured light field to generate its lower resolution version. In addition, we can visually evaluate the performance of the proposed method without downsampling and further increasing the spatial resolution of the original images. In Figure 4.13, we provide a comparison of bicubic resizing, bicubic interpolation, the LFCNN method [4], the DRRN method [67], and the proposed LFSR method. The spatial resolution of each perspective image is increased from 374x540 to 748x1080. The results of the proposed method seem to be preferable over the others with less artifacts. The LFCNN results in sharp images but has some visible artifacts. The DRNN method seems to distort some texture, especially visible in the second example image, while the proposed method preserves the texture well.

(a) Ground Truth | (b) Bicubic resizing(*imresize*) / 25.99 dB | (c) Bicubic interp. / 28.15 dB | (d) LFCNN [4]/ 24.09 dB | (e) DRRN [67]/ 33.65 dB | (f) LFSR / 29.75 dB

Figure 4.11: Visual comparison of different methods. (The worst result image from the dataset is shown here).

(a) 15.03 dB (b) 23.54 dB (c) 23.88 dB (d) 20.08 dB (e) 27.80 dB (f) 33.08 dB (g) 31.55 dB (h) Ground Truth

Figure 4.12: Visual comparison of the novel view. This "Leaves" picture from Kalantari *et al.*[3] paper contains very thin structure and significant amount of occluded regions, which makes it difficult to synthesis novel views. Our architecture produces resonably better result as compared to the state-of-the-art methods. (a) Wanner *et al.* [6]. (b) Tao *et al.* [7]. (c) Wang *et al.* [8]. (d) Jeon *et al.* [9]. (e) Kalantari *et al.* [3]. (f) Proposed Angular SR network. (g) LFSR.

(a) (b) (c) (d) (e)

(f) Bicubic resizing(*imresize*) (g) Bicubic interp. (h) LFCNN [4] (i) DRRN [67] (j) Proposed LFSR

Figure 4.13: Visual comparison of different methods.

Chapter 5

Discussion And Conclusion

In this thesis, we presented a convolutional neural network based light field super-resolution technique. The proposed method consists of two separate convolutional neural networks trained through supervised learning. The architecture of these networks are composed of only three layers, reducing computational complexity. The proposed method shows significant improvement both quantitatively and visually over the baseline bicubic interpolation and another deep learning based light field super-resolution method. In addition, we compared the angular resolution enhancement part of our method against two methods for novel view synthesis. We also demonstrated that enhanced light field results in more accurate depth map estimation due to the increase in angular resolution.

The spatial super-resolution network is designed to generate one perspective image. One may suggest to generate all perspectives in a single run; however, this would result in a larger network, requiring larger size dataset and more training. Instead, we preferred to have a simple, specialized, and effective architecture.

Similar to other neural network based super-resolution techniques, the method is designed to increase the resolution by an integer factor (two). It can be applied multiple times to increase the resolution by factors of two. A non-integer factor size change is also possible by first interpolating using the proposed method and

then downsampling using a standard technique.

The network parameters are optimized for a specific light field camera. For different cameras, the specific network parameters, such as filter dimensions, may need to be optimized. We, however, believe that the overall architecture is generic and would work well with any light field imaging system once optimized.

Bibliography

[1] 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 Trans. on Graphics (TOG)*, vol. 24, no. 3, 2005, pp. 765–776.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. on Graphics (Proceedings of SIGGRAPH Asia)*, vol. 35, no. 6, 2016.

[4] 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 *Proc. of the IEEE Int. Conf. on Computer Vision Workshops*, 2015, pp. 24–32.

[5] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, 2017.

[6] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 41–48.

[7] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *IEEE Int. Conf. on Computer Vision (ICCV)*, December 2013.

[8] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *IEEE Int. Conf. on Computer Vision*, 2015, pp. 3487–3495.

[9] 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 *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1547–1555.

[10] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. on Graphics (TOG)*, vol. 26, no. 3, p. 69, 2007.

[11] A. Wang, P. R. Gill, and A. Molnar, "An angle-sensitive cmos imager for single-sensor 3d photography," in *IEEE Int. Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 2011, pp. 412–414.

[12] R. Ng, "Fourier slice photography," in *ACM Trans. on Graphics (TOG)*, vol. 24, no. 3, 2005, pp. 735–744.

[13] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *IEEE Int. Conf. on Computational Photography*, 2009, pp. 1–8.

[14] "Lytro, inc." https://www.lytro.com.

[15] "Raytrix, gmbh." https://www.raytrix.de.

[16] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1027–1034.

[17] V. Boominathan, K. Mitra, and A. Veeraraghavan, "Improving resolution and depth-of-field of light field cameras using a hybrid imaging system," in *IEEE Int. Conf. on Computational Photography*. IEEE, 2014, pp. 1–10.

[18] X. Wang, L. Li, and G. Hou, "High-resolution light field reconstruction using a hybrid imaging system," *Applied Optics*, vol. 55, no. 10, pp. 2580–2593, 2016.

[19] M. Z. Alam and B. K. Gunturk, "Hybrid stereo imaging including a light field and a regular camera," in *Signal Processing and Communication Application Conference (SIU)*. IEEE, 2016, pp. 1293–1296.

[20] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *IEEE Int. Conf. on Computational Photography*, 2009, pp. 1–9.

[21] S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4d light fields," in *European Conf. on Computer Vision*. Springer, 2012, pp. 608–621.

[22] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2013, pp. 3280–3287.

[23] M. S. K. Gul and B. K. Gunturk, "Spatial and angular resolution enhancement of light fields using convolutional neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2146–2159, May 2018.

[24] G. Lippmann, "Epreuves reversibles, photographies integrales," *J. Academie des sciences*, pp. 446–451, 1908.

[25] A. Gershun, "The light field," *J. of Mathematics and Physics*, vol. 18, 1936.

[26] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology*, 1991.

[27] E. H. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 99–106, 1992.

[28] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. of the 23rd Annual Conf. on Computer Graphics and Interactive Techniques*. ACM, 1996, pp. 31–42.

[29] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. of the 23rd Annual Conf. on Computer Graphics and Interactive Techniques*. ACM, 1996, pp. 43–54.

[30] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. of SIGGRAPH*, 2000, pp. 297–306.

[31] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 924–934, 2006.

[32] C. Perwass and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," in *Proc. SPIE Human Vision and Electronic Imaging*, 2012, p. 829108.

[33] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 39–46.

[34] J. C. Yang, M. Everett, C. Buehler, and L. McMillan, "A real-time distributed light field camera." *Rendering Techniques*, vol. 2002, pp. 77–86, 2002.

[35] J. Unger, A. Wenger, T. Hawkins, A. Gardner, and P. Debevec, "Capturing and rendering with incident light fields," UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES, Tech. Rep., 2003.

[36] A. Manakov, J. Restrepo, O. Klehm, R. Hegedus, E. Eisemann, H.-P. Seidel, and I. Ihrke, "A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 47–1, 2013.

[37] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The J. of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[38] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[39] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 22–28.

[40] J. Wu, H. Wang, X. Wang, and Y. Zhang, "A novel light field super-resolution framework based on hybrid imaging system," in *Visual Communications and Image Processing (VCIP)*. IEEE, 2015, pp. 1–4.

[41] B. K. G. M Umair Mukati, "Hybrid-sensor high-resolution light field imaging," in *Signal Processing and Communication Application Conference (SIU), 2017 25th*. IEEE, 2017.

[42] M. Z. Alam and B. K. Gunturk, "Hybrid light field imaging for improved spatial resolution and depth range," *Machine Vision and Applications*, pp. 1–12, 2018.

[43] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1831–1838.

[44] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM Trans. on Graphics (TOG)*, vol. 34, no. 1, p. 12, 2014.

[45] D. C. Schedl, C. Birklbauer, and O. Bimber, "Directional super-resolution by means of coded sampling and guided upsampling," in *IEEE Int. Conf. on Computational Photography (ICCP)*. IEEE, 2015, pp. 1–10.

[46] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 606–619, 2014.

[47] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2015, pp. 576–584.

[48] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 769–777.

[49] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1439–1451, 2016.

[50] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[51] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2013, pp. 633–640.

[52] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[53] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776.

[54] L. Xu, J. S. Ren, Q. Yan, R. Liao, and J. Jia, "Deep edge-aware filters." in *ICML*, 2015, pp. 1669–1678.

[55] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conf. on Computer Vision*. Springer, 2016, pp. 649–666.

[56] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search," *Biomedical Optics Express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[57] L. Fang, S. Li, R. P. McNabb, Q. Nie, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Fast acquisition and reconstruction of optical coherence tomography images via sparse representation," *IEEE transactions on medical imaging*, vol. 32, no. 11, pp. 2034–2049, 2013.

[58] L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 407–421, 2017.

[59] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conf. on Computer Vision*. Springer, 2014, pp. 184–199.

[60] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[61] ——, "Deeply-recursive convolutional network for image super-resolution," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[62] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conf. on Computer Vision*. Springer, 2016, pp. 694–711.

[63] A. S. Raj, M. Lowney, and R. Shah, "Light-field database creation and depth estimation," https://lightfields.standford.edu.

[64] 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 *Proc. of the 22nd ACM Int. Conf. on Multimedia*, 2014, pp. 675–678.

[65] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.

[66] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields." in *VMV*, 2013, pp. 225–226.

[67] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

SUPER RESOLUTION OF LIGHT FIELDS USING CONVOLUTIONAL NEURAL NETWORK

ORIGINALITY REPORT

19%	9%	16%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Nashville State Community College Student Paper	1%
2	Nima Khademi Kalantari, Ting-Chun Wang, Ravi Ramamoorthi. "Learning-based view synthesis for light field cameras", ACM Transactions on Graphics, 2016 Publication	1%
3	"Computer Vision – ECCV 2016", Springer Nature, 2016 Publication	1%
4	www.umairmukati.com Internet Source	1%
5	M. Zeshan Alam, Bahadir K. Gunturk. "Hybrid light field imaging for improved spatial resolution and depth range", Machine Vision and Applications, 2017 Publication	1%
6	Submitted to Higher Education Commission	