Light Field Super Resolution with Convolutional Neural Networks

by
Andrew Hou

A Thesis submitted in partial fulfillment of the requirements for Honors
in the Department of Applied Mathematics and Computer Science at Brown University

Providence, Rhode Island
May 2018

This thesis by Andrew Hou is accepted in its present form by
the Department of Applied Mathematics and Computer Science as satisfying the research requirement
for the awardment of Honors.

Date ______________ ______________________________
Jerome Darbon, Reader

Date ______________ ______________________________
James Tompkin, Reader

Abstract

Light field reconstruction given a sparsely sampled set of light field views yields various artifacts including ghosting and tearing due to information asymmetry between the spatial and angular domain. In particular, the angular detail of the light field is damaged by undersampling. To balance spatial and angular information, we model the problem of light field super resolution as an angular detail restoration problem on 2D epipolar plane images (EPIs). We take advantage of the unique structure of the EPI and utilize the CNN-based model from Wu et al. [1] that performs angular super resolution. We modify their pipeline by removing the blur-deblur steps and maintaining the standalone CNN architecture with some small modifications. We argue that the blur and deblur steps are unnecessary and do not improve the quality of the reconstruction or the visual coherence of the super resolved light field. We compute the peak signal to noise ratio (PSNR) values between our reconstructions and a groundtruth light field to show that our model achieves high performance even compared to current state of the art methods for light field super resolution.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Light fields are commonly used to represent and reconstruct 3D images of a scene. They are often captured using multi-camera arrays, and more recently by commercial light field cameras like the Lytro light field camera. Due to limited sensor resolution, there is often a tradeoff between spatial and angular resolution of the light field and as a result the light field can lose angular detail due to undersampling. This poses the problem of restoring angular detail for the light field given an initial sparsely sampled light field such that we reduce the information asymmetry between the spatial and angular domains. We refer to this problem as Light Field Super Resolution as our model essentially performs angular super resolution while preserving high spatial resolution.

We frame this problem as a CNN-based, angular detail restoration problem on 2D epipolar plane images (EPIs) sampled from 2D light field images that each represent a different view of the scene. We take advantage of the EPI's unique and geometrically simple structure and argue that our CNN model can more easily learn the patterns observed in EPIs than in full light field images.

We reconstruct a more densely sampled light field with higher angular resolution using a 3 step reconstruction technique that we discuss in Chapter 5. We report peak signal to noise ratio (PSNR) values between 2D light field views generated by our model and the corresponding views from the high resolution groundtruth light field. Our PSNR values are close to state-of-the-art methods, and our reconstructed light field views generally do not suffer from unwanted artifacts like ghosting and tearing that result from information asymmetry between the spatial and angular domain.

Chapter 2

Related Work

Light field super resolution techniques can be divided into two main categories based on their nature: those that rely on depth estimation, and those that do not. We discuss the inherent benefits and costs for both depth dependent and depth independent approaches.

2.1 Depth Dependent Super Resolution

There are a number of techniques that approach light field super resolution by using an estimated depth map of the scene. Wanner and Goldluecke [5] generate a depth map and use that to warp the input light field images to the novel views. Kalantari et al. [8] used two sequential CNNs to simultaneously model depth and color by minimizing the error between ground truth images and synthetic views. However, their approach fails to handle occluded regions as they generate ghosting and tearing artifacts. One of the key problems with super resolution approaches that rely on depth is that their performance is entirely dependent on the quality of depth estimation at every point. Therefore, they will inevitably fail where depth estimation is poor: in occluded and transparent regions and for non-Lambertian surfaces.

2.2 Depth Independent Super Resolution

Our framework for modeling light field super resolution as an angular detail restoration problem on EPIs is primarily based off of the work of Wu et al.[1]. They use a 3 step framework that involves first blurring the spatial domain of epipolar plane images (EPIs) to account for information asymmetry between the spatial and angular domains, feeding them into a 3 layer fully convolutional neural network to restore angular detail, and finally deblurring the network generated EPIs along the spatial domain to restore spatial detail. The final deblur operation is a deconvolution algorithm from Krishnan et al. [3]. Our pipeline is based off of [1] with some adjustments to network hyperparameters and architecture as well as how we train the model. We also argue that their spatial blur and deblur steps are unnecessary and do not necessarily produce higher quality light field images than just the standalone CNN. The authors of [1] produce depth maps and novel light field views and report PSNR values for their generated light field images. An alternative network architecture is presented in Gul et al. [2]. As opposed to [1] which trains a network to predict horizontal and vertical adjacent pixels in the EPIs and focuses on enhancing angular resolution, [2] trains two networks on lenslet images, one that enhances angular resolution and another that enhances spatial resolution. The network that enhances spatial resolution predicts horizontal, vertical, and diagonal pixels to enlarge the lenslet image. Their framework reports high PSNR values for a variety of light fields and in general produces sharp images with very few artifacts.

Chapter 3

Epipolar Plane Images

We model the problem of Light Field Super Resolution as an angular detail restoration problem along EPIs. Given a 4D light field $L(x, y, s, t)$, where x and y are the spatial dimensions of the light field and s and t are the angular dimensions, we construct 2D EPIs by fixing y to be some y* and t to be some t*. This can be restated as $E_{y^*,t^*}(x, s)$. We're provided with 4D light fields represented as nxn grids of 2D views. An example of this representation can be observed in Figure 3.1. Intuitively, a 2D EPI is constructed by taking the same horizontal scanline from each image in a row of the nxn light field and stacking them. Figure 3.2 provides a visualization of an EPI generated from the Amethyst Light Field in the Stanford Light Field Dataset.

Each of the slanted lines in an EPI corresponds to a particular object in the scene that the light field represents. The slope of the line has a direct correspondence with the depth of the object in the scene. An EPI line with a gentle slope represents an object with low depth, whereas a steep EPI line represents an object that has high depth. A horizontal EPI line would represent an object at 0 depth in the image, and a vertical line would denote an object that is an infinite distance away.

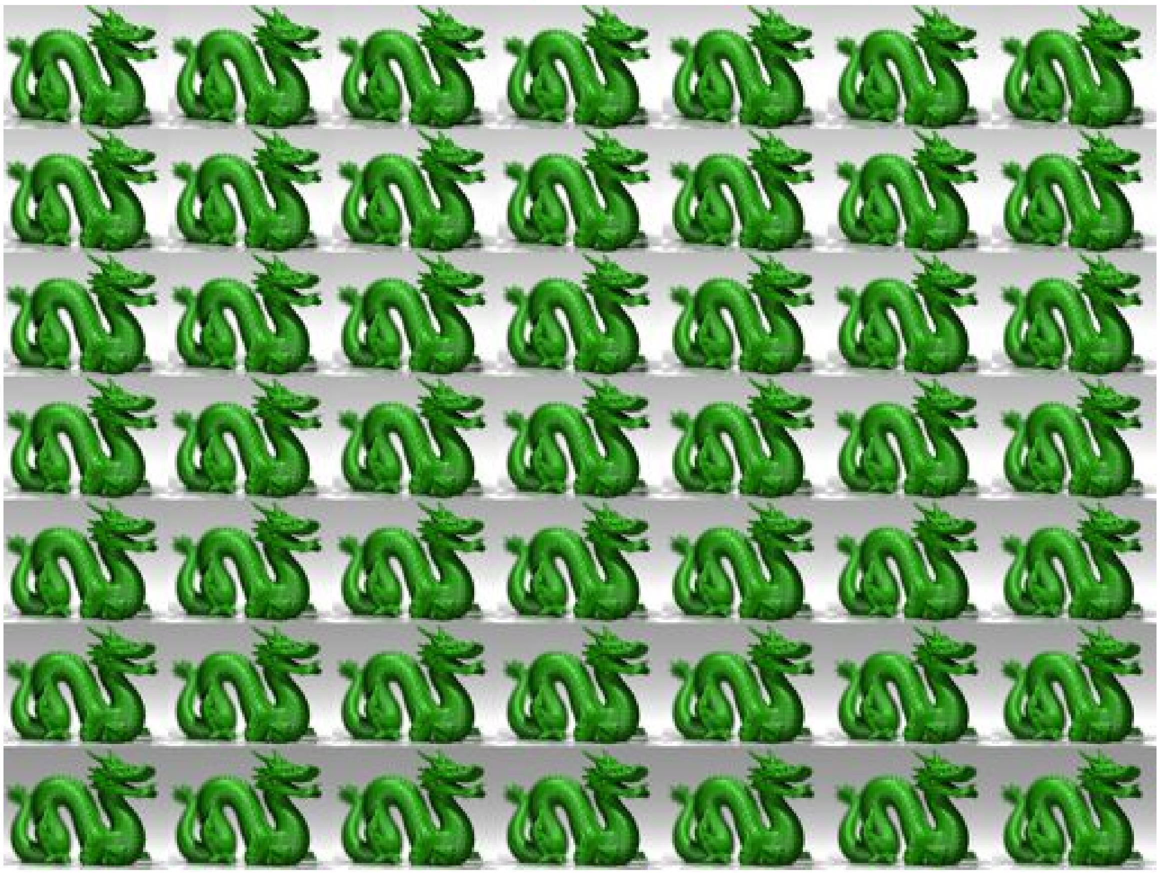

Figure 3.1: 7x7 light field of a green dragon. Each view represents the scene viewed from a different set of angles s and t.

Figure 3.2: A 2D EPI constructed from the 17x17 Amethyst Light Field from the Stanford Light Field Dataset. This particular EPI was created by taking the 453^{rd} row of every image in the first row of 17 images in the Amethyst Light Field and stacking them vertically.

Chapter 4

Super Resolution Model

4.1 Angular Detail Restoration Convolutional Neural Network

We use the convolutional neural network architecture presented in [1] for angular detail restoration on sparsely sampled EPIs. The sparsely sampled EPI is upsampled vertically to the desired angular resolution and converted to Y-Cb-Cr space. We take the Y (luminance) channel of the EPI and split it vertically into 17x17 pixel sub-EPI patches with stride 14 such that each pair of sub-EPIs has an overlap of 3 columns. These upsampled Y channel sub-EPI patches are fed to the CNN, which attempts to learn a residual image between the upsampled input patches and corresponding patches generated from the groundtruth light field at the desired angular resolution. We do not feed the Cb and Cr channels to the network, as these channels are blurrier than the Y channel and are thus less useful in reconstructing a super resolved light field.

The convolutional network is fully convolutional, and consists of 3 convolutional layers. The first layer is a feature extraction layer that contains 64 filters each of size 1x9x9 that perform convolution over 9x9 spatial patches and generate 64 channels (feature maps). The second layer is simply a non linear mapping and consists of 32 filters, each of which is 64x5x5. The third convolutional layer is a detail reconstruction layer that consists of a single filter of size 32x5x5. This third layer produces a residual image, which is summed with the upsampled input EPI to produce the final output EPI with higher angular detail. We use the rectified linear unit (RELU) activation function in our fully convolutional network, and as such the first and second convolutional layers are both followed by a RELU layer. A visualization of the CNN architecture is represented in Figure 4.1.

The CNN model learns a residual image R, which ideally is the difference between the upsampled input EPI E and the groundtruth EPI with higher angular detail E_G. That is, we want the network to learn:

$$R = E_G - E$$

The loss function L that we aim to minimize is the mean square error between the groundtruth high resolution EPI E_G and the output EPI that our CNN model produces $f(E)$. This can be restated as:

$$L = \frac{1}{2} * ||E_G - f(E)||^2$$

Since our network is a residual network, we reformulate this into the mean square error between all groundtruth residual images and residuals predicted by our network which is an equivalent loss function L_R. This loss function can be represented as:

$$L_R = \frac{1}{n} \sum_{i=1}^{n} ||R^{(i)} - r(E^{(i)})||^2$$

,

where n ranges across all training EPIs, $R^{(i)}$ is the ideal groundtruth residual given the i^{th} groundtruth EPI and its corresponding upsampled input EPI, and $r(E^{(i)})$ is the residual that the network produces.

4.2 Data Preprocessing and Augmentation

To improve training stability and avoid overfitting, we perform data augmentation on the training set. During every epoch for each sub-EPI, we perform a horizontal flip with 0.5 probability; otherwise, we keep the EPI in its original orientation. We also shuffle the training set before every epoch to aid with convergence and to prevent the model from overfitting on the training data. Sub-EPIs fed to the network are normalized between -1 and 1, and we find that this preprocessing step substantially improves training stability and reconstruction results.

4.3 Training Detail

We train our model using the rectified images of light fields from the (New) Stanford Light Field Archive [4]. The Archive contains a collection of 13 light fields with different properties that together form a comprehensive and representative training set. This includes light fields with complex geometry and a wide range of depths. The variety of light fields in the archive results in many different EPI structures and produces a more general and robust model. Each of these light fields contains 289 views arranged in a 17x17 grid, and we generate horizontal EPIs from each light field for training using the horizontal scanlines (rows) of the light field images. In particular, for each row of images in a 17x17 light field, we generate an EPI for every scanline in the light field image by vertically stacking the same scanline from every other image in the row of 17 images. This produces

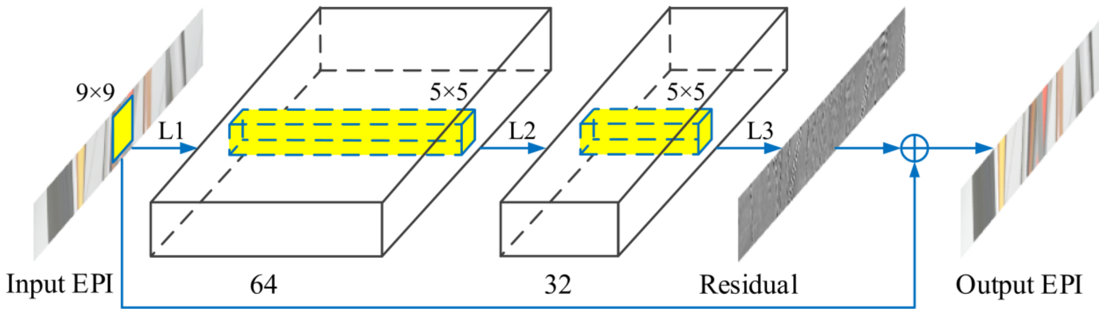

Figure 4.1: The angular detail restoration network (from Wu et al. [1]). It consists of 3 convolutional layers with the first and second layer followed by a RELU layer. The third layer produces a residual image which is summed with the upsampled input EPI to produce the final output EPI with the desired angular resolution and higher angular detail.

an EPI that has 9 rows representing 9 angular views. We upsample this EPI to have 17 rows, the desired angular resolution. The intermediate rows of the upsampled EPI are estimated using bilinear interpolation.

We generate 3 million sub-EPI training examples by repeatedly sampling a random upsampled EPI from a random light field. We divide each selected EPI into 17x17 sub-EPI patches and add these patches to the training set. This sampling process continues until we generate at least 3 million training examples. We experimentally found that 3 million was a sufficient number of examples by evaluating model performance given different numbers of training examples. The model performance starts to stabilize once you train with at least 2 million sub-EPI training examples.

The network is trained for 50 epochs (full cycles through the training set) using a batch size of 64. We use the Adam optimization algorithm for weight updates with a learning rate of 0.0001. The weights in the convolutional layers of the network are initialized by sampling from a Gaussian distribution with a mean of 0 and a standard deviation of 0.001.

We implement the angular detail restoration CNN in Tensorflow-GPU version 1.2.1 and we train the model on an NVIDIA GeForce GTX 1080. Training the model requires approximately 3 hours. This includes sampling, augmenting, and normalizing all of the training data from the Stanford Light Field Archive as well as 50 epochs of training.

Chapter 5

Light Field Reconstruction

5.1 3 Step Reconstruction

To reconstruct a full $n'xn'$ light field given an initial sparsely sampled light field that is nxn, we use a 3 step reconstruction process.

First, we generate our EPIs $E_{y^*,t^*}(x,s)$ constructed from the horizontal scanlines of a row of images in the sparse input light field. We upsample these EPIs along the angular domain to n' views and feed these EPIs through our model to generate our super resolved, high angular detail EPIs, which can then be used to construct intermediate views in the rows of the light field. This process is explained in detail in section 5.2.

Second, we generate our EPIs $E_{x^*,s^*}(y,t)$ constructed from the vertical scanlines of a column of images in the input light field. Our model generates super resolved EPIs, which are then used to create intermediate views in the columns of the light field.

Finally, to complete the missing views in the super resolved light field, we can either use the horizontal EPIs $E_{y^*,t^*}(x,s)$ from our novel views in each row or the vertical EPIs $E_{x^*,s^*}(y,t)$ from our novel views in each column. These are fed to the model, which produces angular super resolved EPIs, which can then be used to construct the remaining novel views in the high resolution light field. This reconstruction technique is visualized in Figure 5.1.

5.2 Reconstructing Light Field Views From EPIs

Our angular detail restoration CNN produces super resolved EPIs at the same spatial and angular resolution as the desired final light field. In our work, we generate sparse 9x9 light fields and super resolve them to be 17x17. The network thus produces horizontal EPIs with 17 rows, or vertical EPIs with 17 columns, depending on which step of the reconstruction we're executing.

Beginning with the first step in the reconstruction, rows of the output EPI indexed by $2*(i-1)+1$,

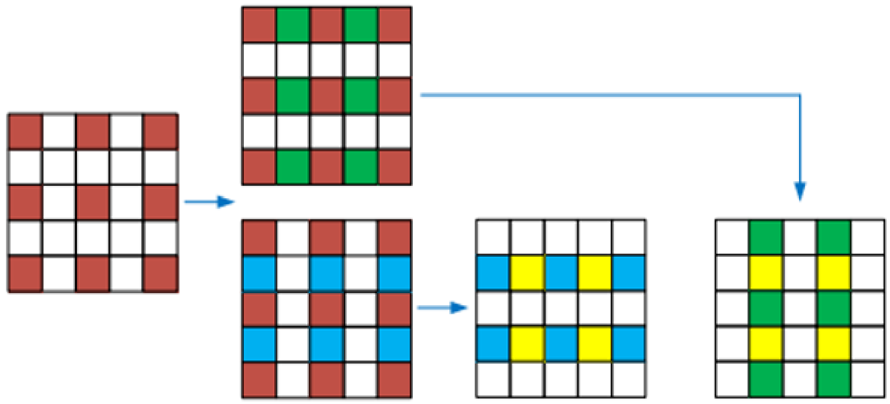

Figure 5.1: A visualization of the reconstruction process. During training, we simulate an undersampled light field by reducing the angular resolution in both s and t by a factor of 2 (a 5x5 light field becomes a 3x3 light field, represented by the red boxes). For each row of light field images, we generate 2D EPIs from the horizontal scanlines and upsample them to the desired angular resolution (5 views). The model learns the missing angular detail, which is then used to construct novel views in each row (green boxes). We repeat this using vertical EPIs in every column of light field images to generate novel views in every column (blue boxes). The final step is to estimate the missing views (yellow boxes) using either horizontal or vertical EPIs from our novel views generated in the first two steps. (from Wu et al. [1])

where $i = 1, 2, ..., 9$ were sampled from the sparse light field and are not needed for reconstruction since those views are provided. All other rows in the EPI are used to reconstruct novel intermediate views in the rows of the super resolved light field. The j^{th} row in a horizontal EPI that was created by stacking the k^{th} horizontal scanline from each image in row m of the light field becomes the k^{th} horizontal scanline in the j^{th} image along row m of the light field. For example, if the EPI was constructed from stacking the 400^{th} row of each image in the first row of images in the 9x9 input light field, 2^{nd}, 4^{th}, 6^{th}, ..., and 16^{th} rows of the corresponding super resolved EPI will be used to reconstruct the 400^{th} horizontal scanline in the 2^{nd}, 4^{th}, 6^{th}, ..., and 16^{th} light field images respectively in the first row of the super resolved 17x17 light field.

The same idea applies for the second step, except that we use the 2^{nd}, 4^{th}, 6^{th}, ..., and 16^{th} columns of the super resolved vertical EPI to reconstruct the 2^{nd}, 4^{th}, 6^{th}, ..., and 16^{th} light field images in a column of the light field.

The third step of the reconstruction handles the remaining views in the light field that need to be estimated. We use either the super resolved horizontal EPIs generated from the novel views produced is the second step, or the vertical EPIs generated from the novel views produced in the first step of the reconstruction.

The number of super resolved horizontal and vertical EPIs we produce depends on the spatial dimensions x, y of the light field. For every row of images in the light field, we construct an EPI for every horizontal scanline in the image. For every column of images in the light field, we construct an EPI for every vertical scanline in the image. Since this is done for each row and column of images in the light field, assuming we have a sparse 9x9 input light field consisting of 2D light field views of height 800 pixels and width 1400 pixels, we would generate 9*800=7200 horizontal EPIs and 9*1400=12,600 vertical EPIs. This ensures that at inference time, for every row of every novel light field view that our framework learns in step 1 of the reconstruction procedure, we have a super resolved horizontal EPI that maps to that row. The same applies for the second step: we have a super resolved vertical EPI for every column of every novel view that the model learns in step 2. In the third step, we generate either horizontal or vertical EPIs equal to the number of rows or columns in the light field image respectively. We do this for each row or column of images that contains novel views, depending on which type of EPI we choose to super resolve to learn the remaining views in the super resolved light field.

Chapter 6

Experiments and Evaluation

6.1 Reconstruction Results

We evaluate our model on the microscope light fields in the Stanford Light Field Archive. In particular, we reconstruct a 16x16 light field given an 8x8 input for the Neurons 20x light field. This light field displays a Golgi-stained slice of a rat's brain and contains complex occlusions, making it a challenging light field for evaluation. We report our PSNR for the reconstructed view in row 9, column 6 of the super resolved 16x16 light field and compare our result with other state of the art techniques for super resolution. See Table 6.1.

6.2 Evaluation

Our angular detail restoration CNN model produces visually plausible novel light field images, as can be seen for the Neurons 20x and Chess light fields from the Stanford Light Field Archive. Our model outperforms all state of the art techniques with the exception of Wu et al. [1], and we manage to achieve similar PSNR values despite not using their blur-deblur steps.
As seen in Figure 6.4, Wang et al. [6] produce very blurry light field images, a result of inaccurate

	Neurons20x (dB)
Wang et al. [6]	17.45
Jeon et al. [7]	23.02
Kalantari et al. [8]	20.94
Wu et al. [1]	29.34
Our model	28.89

Table 6.1: Comparison of PSNR values for a single reconstructed view in the Neurons 20x light field

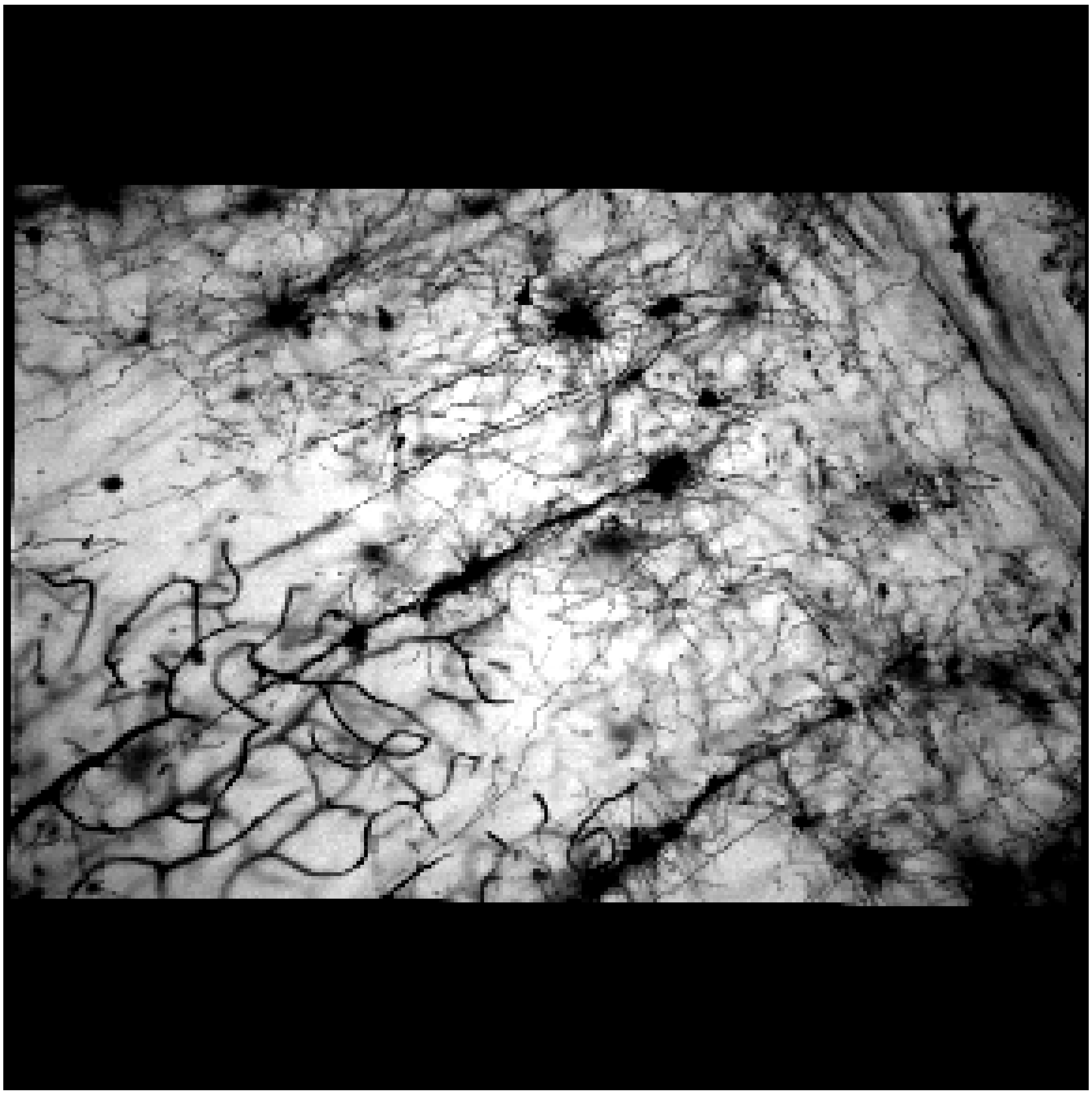

Figure 6.1: Novel light field view estimated from bilinear interpolation for the Neurons20x light field (Row 9, Column 6 in the reconstructed 16x16 light field given an 8x8 input light field), PSNR: 30.01dB

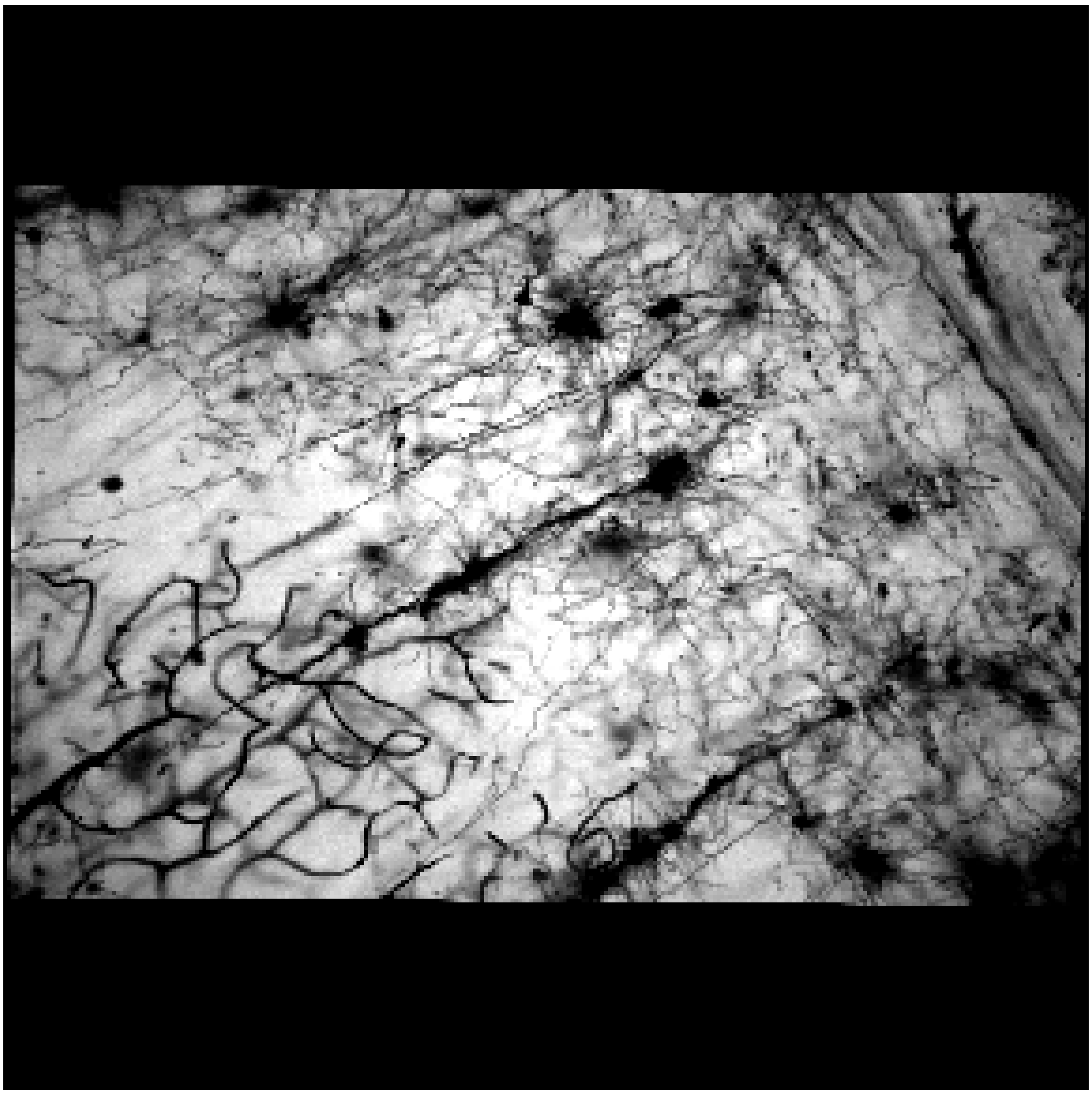

Figure 6.2: Reconstruction of novel light field view for the Neurons20x light field (Row 9, Column 6 in the reconstructed 16x16 light field given an 8x8 input light field), PSNR: 28.89dB

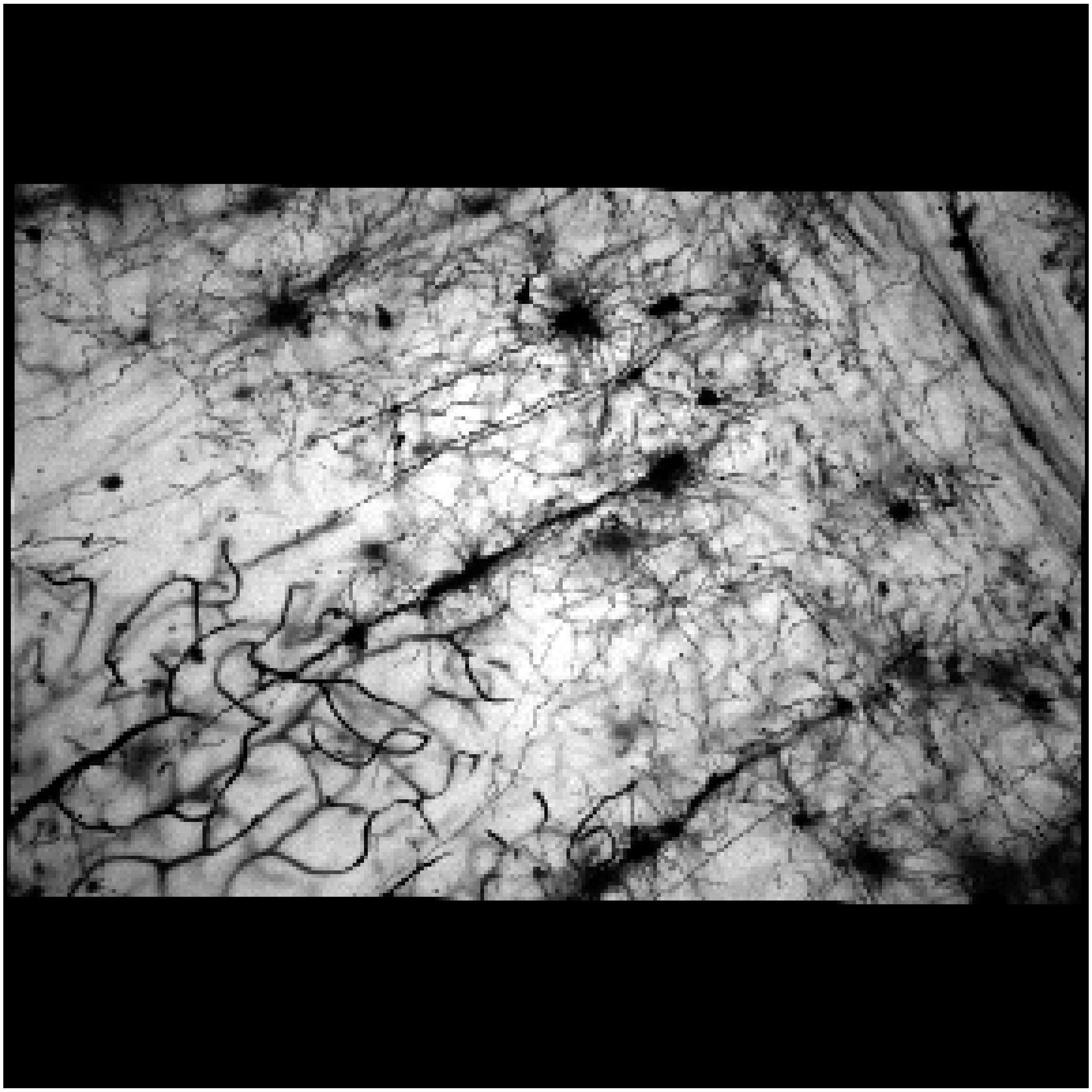

Figure 6.3: Groundtruth light field view at Row 9, Column 6 of the 16x16 Neurons 20x light field

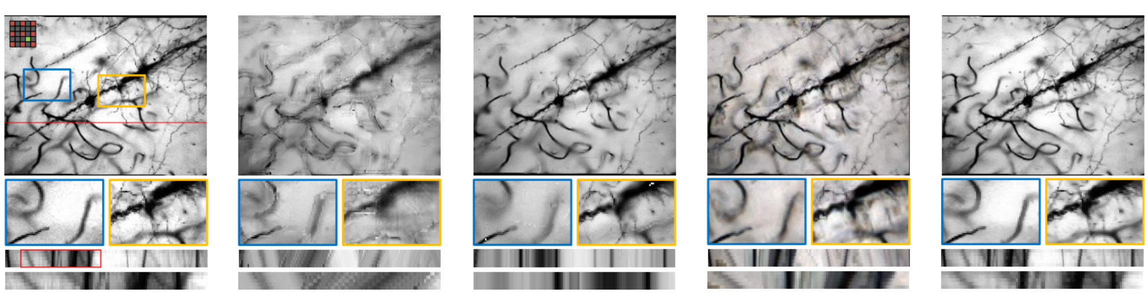

Figure 6.4: Reconstructed views of state of the art methods on the Neurons 40x light field, from left to right: Groundtruth, Wang et al. [6], Jeon et al. [7], Kalantari et al. [8], Wu et al. [1]

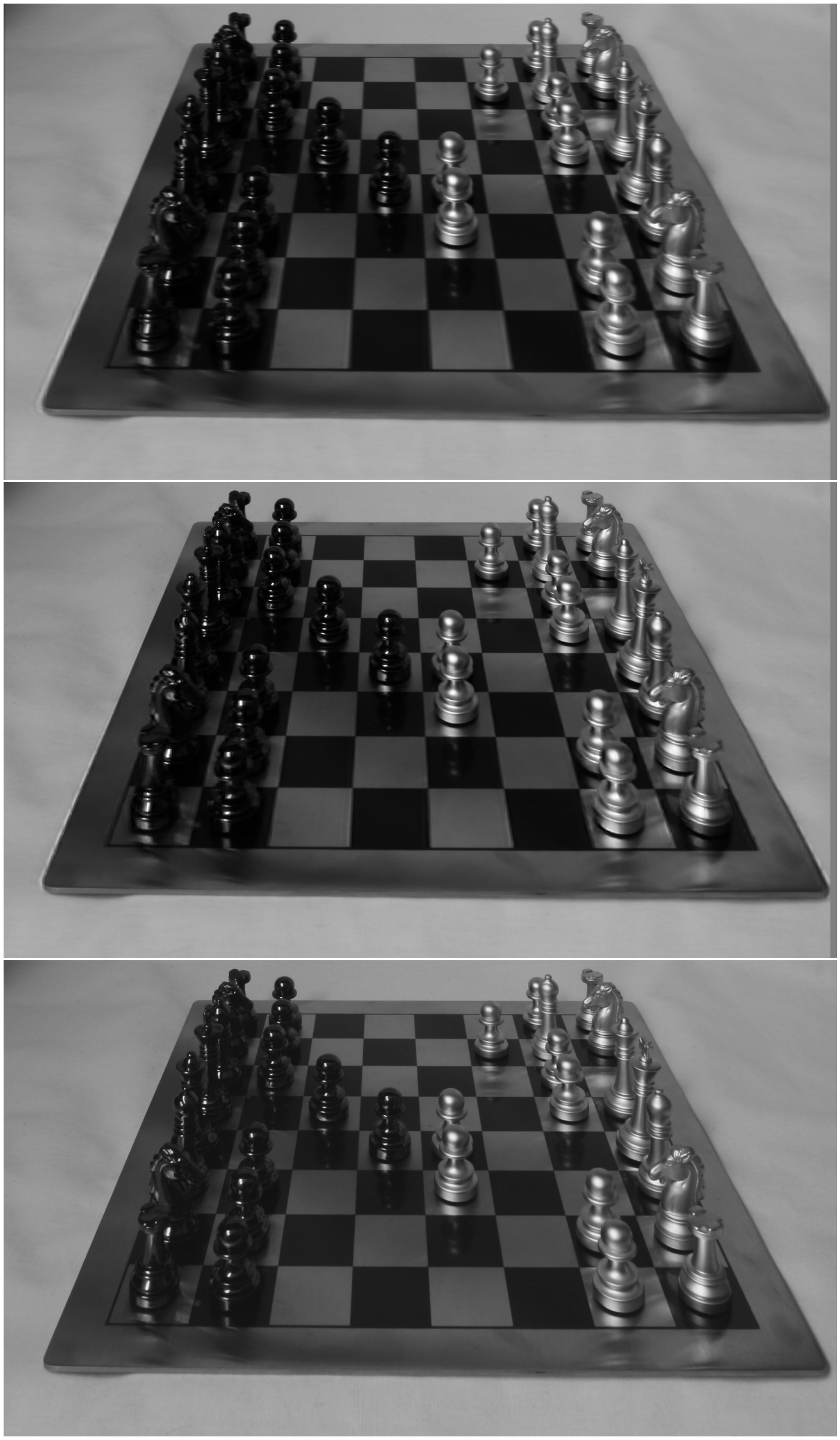

Figure 6.5: Top: Novel view estimated by bilinear interpolation for the 17x17 Chess light field given a 9x9 input (Row 7, Column 2), PSNR: 35.56 dB. Middle: Reconstructed novel view using our CNN model, PSNR: 35.96 dB. Bottom: Groundtruth view. Notice the removal of ghosting artifacts around the chess pieces between the bilinearly interpolated view and our reconstructed view. All images are Y channels in Y-Cb-Cr space.

depth estimations. Jeon et al. [7] are able to achieve higher PSNR but still fail to accurately estimate the depth in the scene. Kalantari et al. [8] fail to properly handle occluded regions in the image, as their occluded regions contain tearing artifacts and blur. Wu et al. [1] is able to quite accurately reproduce the correct depth in the scene and contains very few artifacts in the reconstruction. Our model produces a similar level of visual coherence, as we accurately reproduce the correct depth in different regions of the image (compared with the groundtruth) and are able to handle occluded regions without tearing artifacts. The PSNR of our reconstructed views did not generally improve significantly over bilinearly interpolated views, but they are noticeably sharper and remove some of the ghosting and tearing artifacts from the bilinearly interpolated views. We attribute the lower PSNR values to errors in luminance estimation by the CNN.

6.3 Spatial Blur and Deblur

We argue that the standalone angular detail restoration CNN is sufficient to perform angular super resolution without using the blur and deblur steps highlighted in Wu et al. [1]. The authors assert that spatial blurring is necessary to avoid information asymmetry during training since the angular resolution is damaged by undersampling. The deblur is then used to restore spatial high frequencies after the CNN restores angular detail. However, we achieve almost the same PSNR (within 0.5 dB) both with and without the blur-deblur added to the pipeline when we evaluate on the Chess Light Field from the Stanford Light Field Archive. We also argue that the light field images produced without the blur-deblur steps are more visually coherent and less blurry. We justify this in Figure 6.6.

We also achieve comparable PSNR values to Wu et al. [1] when we evaluate our model on the Neurons 20x light field, further justifying that the standalone CNN is sufficient.

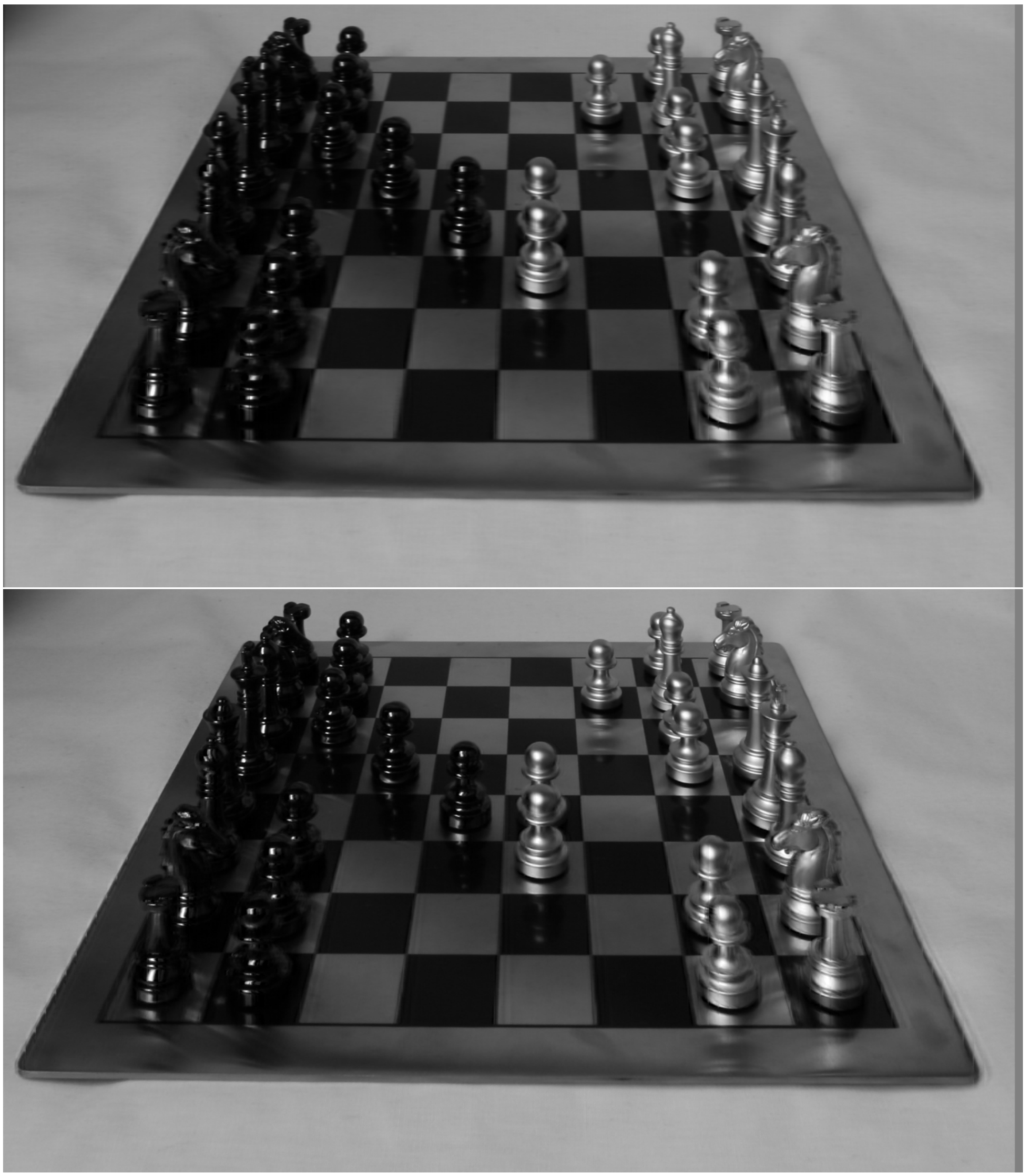

Figure 6.6: Top: A reconstructed novel view for the Chess Light Field using the blur-CNN-deblur pipeline in Wu et al. [1], PSNR: 34.93 dB. Bottom: A reconstructed novel view for the Chess Light Field using just the angular detail restoration CNN with some data augmentation and normalization, PSNR: 35.14 dB. Notice that the bottom image is less blurry (easily seen by comparing the front pawns in the center column of the board and the knights in the back of the board). Both images are Y channels in Y-Cb-Cr space.

Chapter 7

Conclusion

We model the problem of Light Field Super Resolution as a CNN-based angular detail restoration problem on 2D EPIs. We assert that the standalone fully convolutional model without the blur-deblur steps in Wu et al. [1] is sufficient to perform angular super resolution, and we achieve reconstruction results that have comparable PSNR and are more visually coherent than some recent light field super resolution methods. The model is able to accurately estimate depth in reconstructed light field images and improves upon previous methods in reconstructing occluded regions. It is also able to remove artifacts like ghosting and tearing that arise from information asymmetry between the spatial and angular domains. We've shown that the model is capable of performing angular super resolution for general light fields, including challenging light fields like the Stanford Light Field Archive's microscope light fields.

Chapter 8

Future Work

Given the time constraints of this thesis, there is still plenty of room to expand the project. The first line of work would be to evaluate the CNN model on the remaining datasets documented in Wu et al.'s [1] experiments, including the 30 scenes, Reflective29, and Occlusion16 datasets. Potential future work includes modifying Wu et al.'s [1] CNN architecture to see if we can achieve further gains in PSNR and visual coherence of reconstructed light field images. Other potential work includes applying this super resolution model to aid in other problems in the light field domain, or simply for improved depth estimation as reported in Wu et al. [1].

References

[1] Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu, Light Field Reconstruction Using Deep Convolutional Network on EPI, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2017

[2] M. Shahzeb Khan Gul and Bahadir K. Gunturk, Spatial and Angular Resolution Enhancement of Light Fields Using Convolutional Neural Networks, CoRR, 2017

[3] Krishnan, Dilip and Fergus, Rob, Fast Image Deconvolution Using hyper-Laplacian Priors, *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 2009

[4] (New) Stanford Light Field Archive. http://lightfield.stanford.edu/lfs.html.

[5] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. IEEE TPAMI, 36(3):606619, 2014.

[6] T.-C. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion aware depth estimation using light-field cameras. In ICCV, pages 34873495, 2015.

[7] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In CVPR, pages 15471555, 2015.

[8] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG), 35(6), 2016.