12-1-2023

Shift Variant Image Deconvolution using Deep Learning

Arnab Ghosh
ag3671@rit.edu

Shift Variant Image Deconvolution using Deep Learning

by

Arnab Ghosh

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Chester F. Carlson Center for Imaging Science
College of Science
Rochester Institute of Technology

12-01-23

Signature of the Author _____

Accepted by _____
Coordinator, M.S. Degree Program Date

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

M.S. DEGREE THESIS

The M.S. Degree Thesis of Arnab Ghosh
has been examined and approved by the
thesis committee as satisfactory for the
thesis required for the
M.S. degree in Imaging Science

Dr. Grover Swartzlander, Thesis Advisor

Dr. Dimah Dera, Committee

Dr. Carl Salvaggio, Committee

THESIS RELEASE PERMISSION
ROCHESTER INSTITUTE OF TECHNOLOGY
COLLEGE OF SCIENCE
CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

Title of Thesis:

Shift Variant Image Deconvolution using Deep Learning

I, Arnab Ghosh, hereby grant permission to Wallace Memorial Library of R.I.T. to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

Signature _____

Date

Abstract

Image Deconvolution is a well-studied problem that seeks to restore the original sharp image from a blurry image formed in the imaging system. The Point Spread function(PSF) of a particular system can be used to infer the original sharp image given the blurred image. However, such a problem is usually simplified by making the shift-invariant assumption over the Field of View (FOV).

Realistic systems are shift-variant; the optical system's point spread function depends on the position of the object point from the principal axis. For example, asymmetrical lenses can cause space variant aberration.

In this paper, we first simulate our shift-variant aberrations by generating Point Spread Functions using the Seidel Aberration polynomial and use a shift-variant forward blur model to generate our shift-variant blurred image pairs. We then introduce, ShiVaNet. It is a two-stage architecture that builds upon the Learnable Wiener Deconvolution block as described in Yanny, Monakhova, Shuai, and Waller (Yanny et al.) by introducing Simplified Channel Attention and Transpose Attention to improve the performance of the module. We also devise a novel UNet refinement block by fusing a ConvNext-V2 block with Channel Attention and coupling with Transposed Attention Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.). Our model performs better than state-of-the-art restoration models by a factor of 0.2 dB Peak Signal to Noise Ratio.

Acknowledgments

The research presented in this thesis has been made possible through the generous support of both the Chester F. Carlson Center for Imaging Science at the Rochester Institute of Technology and the U.S. Office of Naval Research. It is with deep appreciation that I acknowledge the pivotal role played by these institutions in the realization of this work.

My sincere gratitude extends to my esteemed graduate supervisor, Professor Grover Swartzlander, whose mentorship has not only guided but also profoundly shaped my journey as a researcher. Additionally, I express my thanks to the dedicated members of my committee, Professor Dimah Dera and Professor Carl Salvaggio, for generously offering their time, insights, and invaluable comments, which have significantly enriched the quality of this thesis. I wish to convey my appreciation to Dr. Xiaopeng Peng at RIT, whose collaborative efforts and expertise in Deep Learning have been a source of inspiration and immeasurable benefit to my research. The contributions of other collaborators, including Prateek Srivastava at RIT, as well as Dr. Abbie Watnik, Erin Fleet, and Kyle Novak at the Naval Research Laboratory, have played an integral role in shaping the content and scope of this dissertation. In particular, I extend my special thanks to Professor David Messinger and my dear friends, Touseef Ahmed and Tolga Furkan Aktas, among many others.

In closing, I want to express my deepest appreciation for my late father, Bhaskar Kumar Ghosh, whose encouragement fueled my curiosity for science and whose sacrifices for my well-being have left an indelible mark. To my mother, Mrs. Mukta Ghosh, whose boundless love and enduring patience have been my anchor, and to my sister, Dr. Trisha Ghosh, whose support and encouragement have been a constant source of inspiration—I owe a debt of gratitude for their unwavering belief in my pursuit of scientific passion.

Contents

List of Figures

List of Tables

Acronyms

CNN Convolutional Neural Networks.

FOV Field of View.

GRN Global Response Normalization.

LPIPS Learned Perceptual Image Patch Similarity.

MHSA Multi Headed Self Attention.

MWTA Muti Wiener Transposed Attention.

PSFs Point Spread Function.

PSNR Peak Signal to Noise Ratio.

SCA Simpified Channel Attention.

SSIM Structural Similarity Index Measure.

ViTs Vision Transformers.

Chapter 1

Introduction

Image deconvolution also called Image deblurring is a well-known inverse problem where the true image needs to be approximated given a blurred image. The blurred process can then be defined on the true image $f(x)$ as a Fredholm integral of the first kind :

$$g(x) = \int_{R^2} k(x,u)f(u)du \tag{1.1}$$

The integral kernel $k \in R^2 \times R^2$ is called Point Spread Function (PSFs) of the system. Optical Systems interact with the incoming light from the object and redistribute the intensity of the light causing a blurring effect based on the PSFs.

Image Deconvolution or deblurring is an ill-posed problem that then requires regularization techniques. Conventional image deconvolution algorithms often fall short in robustness, particularly when faced with real-world scenarios involving intricate noise patterns. This limitation stems from their reliance on simplistic noise models, hindering their adaptability to complex and diverse degradation scenarios. One of the most common assumptions is that the kernel operator is shift invariant. The point spread function $k(x,u)$

therefore depends on the object's relative position u concerning x.

$$k(x, u) = k(x - u) \tag{1.2}$$

This simplification allows us to use Fourier transforms to simplify the integral as now the integral transforms into the well-understood convolution integral with a structured PSF.

A lot of classical algorithms have been proposed over the years that assume Shift-Invariance. Wiener deconvolution is one such closed-form algorithm. There are iterative methods such as the Richardson-Lucy algorithm and the fast iterative shrinking-thresholding algorithm Beck and Teboulle (Beck and Teboulle).

1.1 Motivation

A lot of well-built optical systems account for the aberrations by introducing compensatory lenses that seek to mitigate the aberrations. However, the cost increases as hardware costs to correct for aberrations increase dramatically. The cheaper alternative would be to devise a computational solution to the shift variant problem by doing post-processing recovery from the shift variant blur input image. The other major application of this method is on lens protection phase masks that deliberately spread the irradiance to protect the image sensor from oversaturation and damage from damaging sources like high-intensity laser etc. This phase mask goes on top of the optical system and generates a shift variant blur that causes the high-energy beam to spread out more thereby saving the system. Computational image restoration that is sensitive to shift-variance can theoretically allow phase masks to spread high-intensity beams that come obliquely into the system and still restore the image.

1.2 Problem Definition

Image Deconvolution in the shift-variant regime is an extremely ill-posed problem since due to the low pass filter effect of the optical system a lot of frequencies are lost which cannot be recovered if there is no prior information about the world. Many techniques have been developed over the years using the shift-invariant assumption but the shift-variant remains.

Shift-invariant approximation only holds for a small region around the principal axis called the 'paraxial plane'. However, in realistic optical systems, the approximation doesn't hold. Shift variant aberrations like spherical aberration and coma aberration are common in cheaper optical systems like telescopes and cameras. Shift variance makes the integral difficult to solve computationally since the integral is intractable. The usual simplification and computational efficiency that the Fourier transform brings to a shift-invariant system is not possible.

1.3 Summary

The focus of our research is on disregarding shift-invariant simplification and developing a deep-learning-based approach for shift-variant deconvolution. Our paper, first addresses generating shift-variant optical aberration using Seidel polynomials and using a fast method to generate shift-variant blur computationally. We use Seidel aberrations to simulate shift variant blurring which is dependent on the object position and generate the blurring forward model. Now to restore the images we can use image restoration methods like Zhang et al. (2023),Chen, Chu, Zhang, and Sun (Chen et al.),Yanny, Monakhova, Shuai, and Waller (Yanny et al.) which uses deep neural networks to learn the deconvolution function

from the image pairs.Yanny, Monakhova, Shuai, and Waller (Yanny et al.) uses learnable Wiener-deconvolution filters. We build upon this framework to deconvolve our images in a non-blind manner. We use Simplified Channel Attention Chen, Chu, Zhang, and Sun (Chen et al.) and Transposed Attention Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.) to improve the intermediate features generated by the learnable Wiener Deconvolution module. We also define a novel U-Net refinement architecture that is cascaded to the intermediate images generated by our modified Wiener module by introducing ConvNext-V2 Woo, Debnath, Hu, Chen, Liu, Kweon, and Xie (Woo et al.) which improves feature diversity by using Global Response Normalization.

Chapter 2

Image Restoration

2.1 Introduction

Image Restoration is a low-level computer vision problem that has been studied in the image-processing community extensively. One of the challenges in image restoration is the presence of noise in real-world images. Deconvolving noisy images can lead to the amplification of noise, resulting in undesirable artifacts. To address this issue, various regularization techniques are often employed. Regularization methods introduce constraints or penalties during the deconvolution process, helping to strike a balance between image sharpness and noise suppression. The primary purpose of Image-restoration algorithms is to recover the original, clean image from a degraded or noisy version, making it a fundamental problem in low-level computer vision. The issue of noise in real-world images poses a significant challenge in this process, as straightforward deconvolution can inadvertently enhance and propagate noise, introducing unwanted artifacts and reducing the overall quality of the restored image. To mitigate this problem, image restoration methods frequently incorporate regularization techniques. These methods impose constraints or penalties on the optimiza-

tion process during deconvolution, aiming to strike a delicate balance between enhancing image sharpness and suppressing noise. By integrating regularization, image restoration algorithms can effectively trade-off between fidelity to the observed data and adherence to prior knowledge about the expected characteristics of the underlying clean image. This interplay allows for the creation of more robust and visually pleasing restored images in scenarios where noise is a prevalent and challenging factor. Image restoration tasks can be broadly divided into two: 1. Non-Blind Deblurring 2. Blind Deblurring

Non-blind deblurring assumes that the Point Spread Function is already known and the original images are estimated whereas blind deblurring is when the PSF is estimated with the original image as well. Zhang et al. (2022) presents a comprehensive and exhaustive study of image deblurring approaches.

In the next section, we are going to describe some traditional image processing algorithms that tackle the general image deconvolution problem and then provide a primer for image restoration with deep learning. The purpose of this chapter is to familiarize the reader with the traditional image-processing techniques and provide a theoretical background to deep learning theory which allows us to expand and outperform some of the general image-processing techniques. We also discuss image quality metrics that are important to our evaluation of those methods.

2.2 Image Processing Methods

There are various algorithms for image deconvolution, each with its strengths and weaknesses. Here are some commonly used algorithms for image deconvolution:

The Richardson-Lucy algorithm Fish et al. (1995) iteratively refines an estimate of the original image by alternating between a prediction step and an update step. Given a

degraded image and a PSF that describes the blurring, the algorithm attempts to recover the original image. The prediction step involves convolving the current estimate with the PSF to generate a blurred image, and the update step adjusts the estimate based on the ratio of the observed image to the predicted one. This process is repeated over multiple iterations to refine the estimate and improve the deblurring result.

The Wiener deconvolution is a well-known method for non-blind deconvolution. It uses the Wiener filter to estimate the original image by considering the signal-to-noise ratio the power spectra of the degraded image and the PSF. The Wiener deconvolution is effective when the degradation process is known and the noise characteristics are well understood. The Constrained Least Squares deconvolution Ng et al. (2002) is a non-iterative approach that solves a regularized least squares problem to estimate the original image. It introduces a regularization term to control noise amplification during deconvolution. Maximum Likelihood deconvolution is a statistical approach that aims to maximize the likelihood of the observed degraded image given the estimated image and the PSF. It assumes a probabilistic model of the image formation process and optimizes the likelihood function to estimate the original image.

It's worth noting that these are just a few examples of image deconvolution algorithms, and there are many other techniques and variations available. The choice of algorithm depends on factors such as the characteristics of the degraded image, noise level, available information about the blurring process, computational resources, and specific application requirements. We are going to expand on Wiener Deconvolution since it has direct consequences for our research.

2.3 Wiener Deconvolution

Wiener deconvolution is a foundational method in image restoration that leverages statistical insights to address the challenge of noise in observed images. This approach seeks to minimize the mean square error between the estimated image and the true image, taking into account the power spectra of both the true image and the observed image degraded by a blurring operator.

Mathematically, the Wiener deconvolution formula is derived from a statistical estimation perspective, assuming a probabilistic model for both the degradation process and the noise present in the observed image. The primary objective is to find the Least Mean Square error, expressed as $\epsilon(f) = E\left|X(f) - \hat{X}(f)\right|^2$, where $X(f)$ represents the true image in the frequency domain and $\hat{X}(f)$ is the estimated image.

The formulation involves the convolution operation $*$ and the power spectral densities $S(f)$ and $N(f)$ representing the true image and noise, respectively. The expression is given by:

$$\epsilon(f) = [1 - G(f)H(f)][1 - G(f)H(f)]^* S(f) + G(f)G^*(f)N(f) \tag{2.1}$$

Here, $G(f)$ represents the frequency response of the Wiener filter, and $H(f)$ is the frequency response of the blurring operator. The Wiener filter is a critical component in the deconvolution process as it aims to strike a balance between enhancing the details of the true image and suppressing the noise introduced during the imaging process.

To derive the optimal form of the Wiener filter, we calculate the Wirtinger derivative of the mean square error concerning $G(f)$ and set it equal to zero:

$$\frac{d\epsilon(f)}{dG(f)} = 0 \Rightarrow G^*(f)N(f) - H(f)[1 - G(f)H(f)]^* S(f) = 0 \tag{2.2}$$

Rearranging this expression yields the final form of the Wiener filter:

$$G(f) = \frac{H^*(f)S(f)}{|H(f)|^2 S(f) + N(f)} \tag{2.3}$$

The Wiener filter, when applied during the deconvolution process, efficiently mitigates noise amplification, resulting in a more accurate and visually pleasing restoration of the true image.

2.4 Image Restoration based on Deep Neural Networks

2.4.1 Introduction

Traditional methods for image deconvolution, aiming to restore a sharp image degraded by a known blur, relied heavily on mathematical models and signal processing techniques. Wiener filtering employed statistical assumptions about the noise and blur to estimate the original image. Inverse filtering, on the other hand, directly inverted the blur kernel, but was sensitive to noise and often resulted in amplified artifacts. Regularization techniques like Tikhonov regularization were incorporated to control these artifacts by introducing smoothness constraints, but they could lead to over-smoothing and loss of fine details. These traditional methods, despite their historical significance, often required careful parameter tuning and struggled to handle complex blur scenarios, paving the way for more advanced deep learning approaches in recent years. Deep Learning has revolutionized the research area of computer vision by eliminating handcrafted feature engineering that relied on human intuition, making it tougher to generalize on large datasets. Deep Learning allows us to learn hierarchical features directly from the data itself automatically. One of the most successful methods used for images is the Convolutional Neural Network or

CNN. Convolutional neural networks use the basic operation of convolution and learnable features to learn features that can be further used for vision tasks like image segmentation, detection and restoration, etc. Transformers,Vaswani et al. (2017) are a relatively new class of neural networks used in natural language processing to establish long dependencies within a sentence which was later repurposed to find global context in images Dosovitskiy et al. (2021). The fundamental mechanism behind Transformers is Self-Attention. It uses attention pooling to bias selection over values just like how humans pay attention to a particular object only when needed.

Both of these methods are essential to our network and so this section is dedicated to providing a primer to these methodologies as well as a special kind of network architecture called U-Net. In this section, we will discuss the relevant research in more depth to understand the theoretical underpinning of our deep learning model. We will first briefly touch on Artificial Neural Networks, then move on to Convolutional Neural Networks then describe U-Net Architecture. Finally, we will get into Vision Transformers and the primary role of Self-Attention in the network.

2.4.2 Feed-forward Neural Networks

The basic neural network model, inspired by biological neurons, can be mathematically expressed as a linear combination of non-linear basis functions. For a two-layer network, the activations (a_j) of the hidden layer are given by the equation:

$$a_j = \sum_{i=1}^{D} w_{ji}^{1} x_i + w_{j0}^{1}$$

where x_i represents input variables from 1 to D, $j = 1, \ldots, M$, and the superscript (1) indicates the layer number. The parameters w_{ji} are weights, and w_{j0} is the bias of the

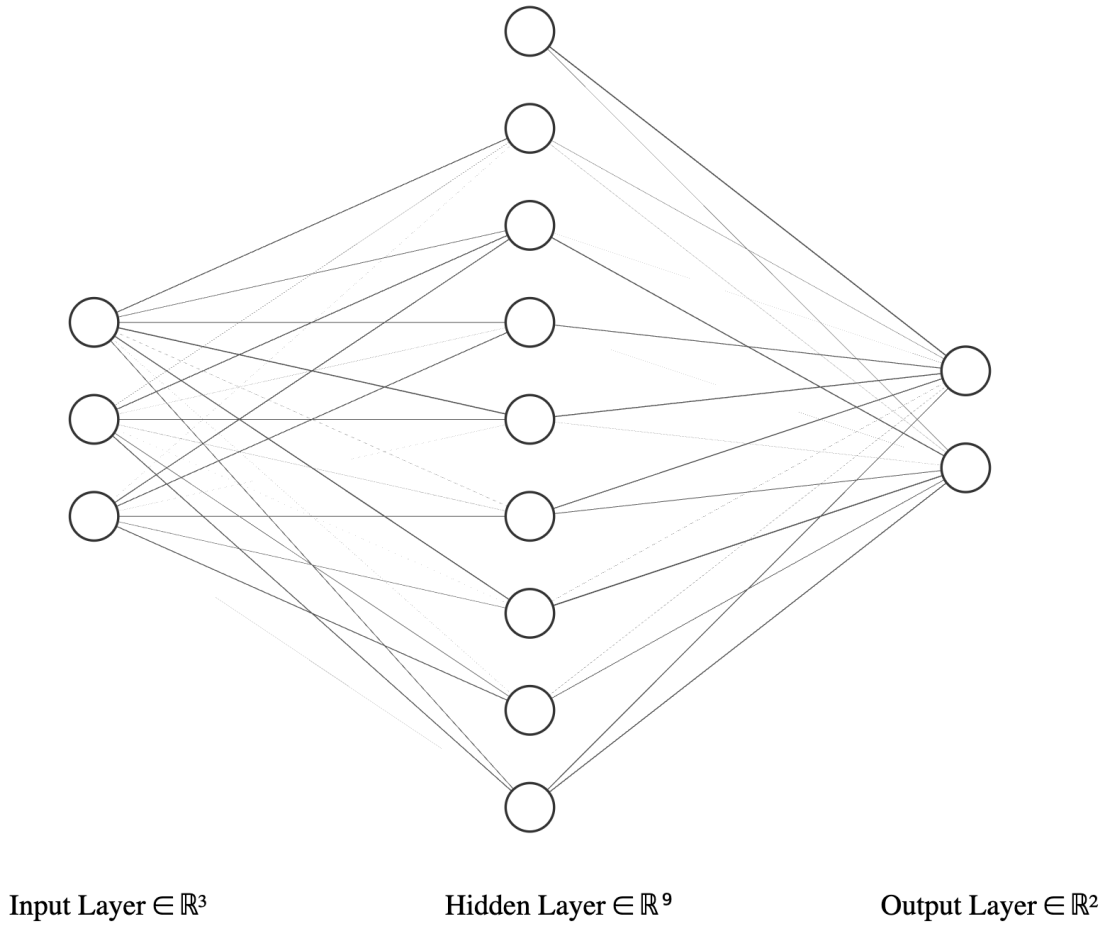

Figure 2.1: Network Diagram for the two-layered neural network

network. A non-linear activation function h then transforms these activations to generate the output. Assuming a sigmoid activation function, a two-layer network can be represented mathematically as:

$$y_k(x, w) = \sigma \left(\sum_{j=1}^{M} w_{kj} h \left(\sum_{i=1}^{D} w_{ji} x_i + w_{j0} \right) + w_{k0} \right)$$

The error function (or loss function) can be defined as the Mean Squared Error between predicted and actual values:

$$\mathcal{L}(w) = \frac{1}{2N} \sum_{k=1}^{N} \|y_k(x, w) - y_k\|^2$$

where N is the number of training examples, y_k is the actual output for the k-th example, and $\|\cdot\|^2$ denotes the L2 norm. Minimizing the L2 norm is equivalent to maximizing the likelihood function of the probability distribution of the traget variable conditioned over the input and weights.

The goal during training is to initialize the network parameters randomly and then minimize this error function using gradient descent. The gradient descent update rule for the weights (w) is given by:

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \alpha \frac{\partial \mathcal{L}(w)}{\partial w_{ij}^{(l)}}$$

where α is the learning rate, and $\frac{\partial \mathcal{L}(w)}{\partial w_{ij}^{(l)}}$ represents the partial derivative of the loss with respect to the weights. This iterative process is performed for multiple epochs until convergence. Bishop (2006) is a good source for having a greater understanding of the concepts underlying neural networks.

2.4.3 Convolutional Neural Networks

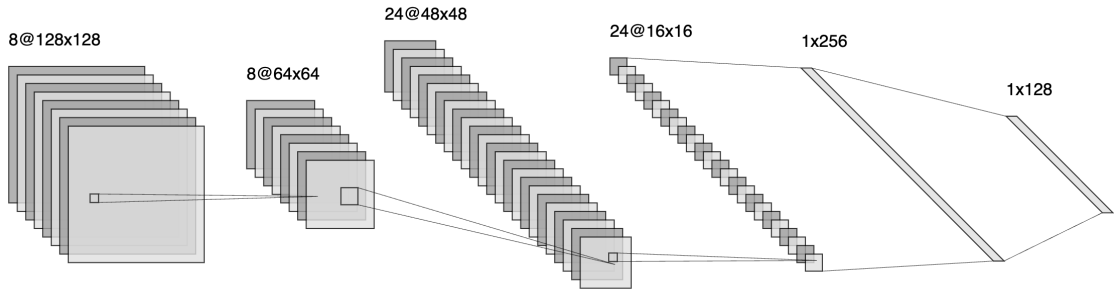

Figure 2.2: Convolutional Neural Network: LeNet Architecture. Learnable filters with variable parameters slide through the image to perform a convolution. Non-Linear activation and Pooling layers outside to reduce feature size.

Convolutional Neural Networks (CNN), also known as ConvNets, are a type of deep learning algorithm specifically designed for processing data with a grid-like structure, such as images and time series data. They have achieved remarkable success in various fields, including computer vision, image recognition, natural language processing, and medical diagnosis.

Convolutional Neural Networks apply filters (also called kernels) to extract features from the input data. The filters slide across the input, performing a dot product operation with the local receptive field at each position. The resulting feature map highlights specific patterns or features in the input. Unlike traditional neural networks where each weight is used once, CNNs utilize shared weights, meaning the same filter is applied across the entire input image. This reduces the number of parameters needed and helps capture local spatial dependencies.

2.4.4 Vision Transformer

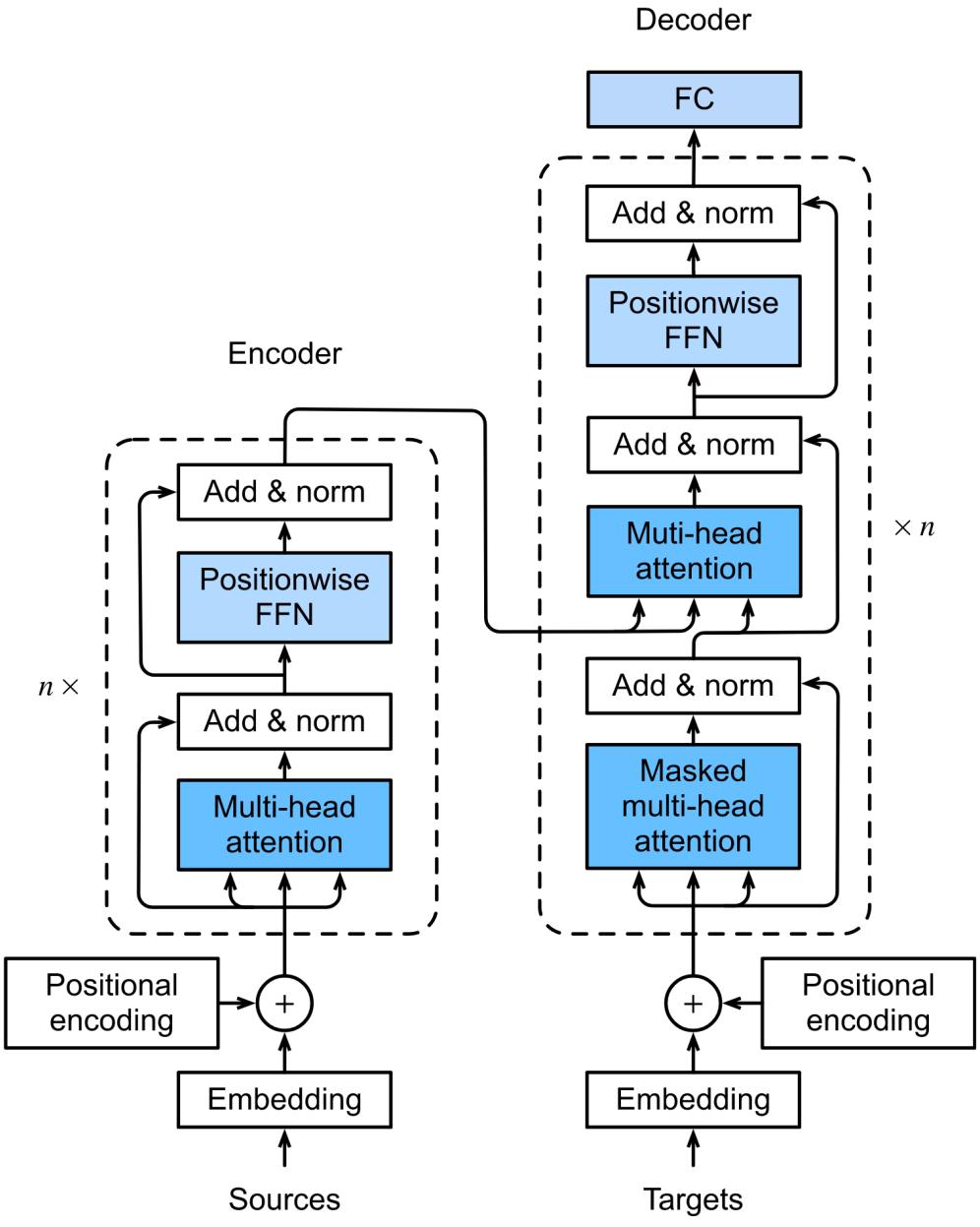

Figure 2.3: The Transformer Architecture: Adapted from Zhang et al. (2023).

Human Visual Systems use nonvolitional cues and volitional cues to focus or deploy attention to a particular object.Vision Transformers (ViTs) have revolutionized the field of computer vision by demonstrating remarkable performance in tasks like object recognition and image segmentation. A key component of ViTs is the multi-head attention mechanism, which allows them to capture complex relationships within the image data.

Traditional attention mechanisms perform a single pooling operation, limiting their ability to capture diverse dependencies. Multi Headed Self Attention (MHSA) addresses this by utilizing parallel attention pooling with different representation subspaces. This enables the model to simultaneously focus on various aspects of the data and capture a wider range of relationships, including both local and global dependencies.

The process involves transforming the original queries, keys, and values, 2.3 into different subspaces using independent linear projections. These transformed representations are then fed into parallel attention-pooling modules, each focusing on specific types of relationships. The outputs of these modules are then concatenated and processed with a final linear projection to produce the final output. This allows the model to combine knowledge from diverse attention behaviors and achieve a richer understanding of the image data.

$$\text{MHSA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.4}$$

Multi-head attention offers several benefits. It captures diverse relationships, improves performance on various tasks, and increases model flexibility by allowing the number of attention heads to be adjusted. This makes it a crucial component of ViTs and contributes significantly to their success in computer vision.

2.4.5 U-Net Architecture

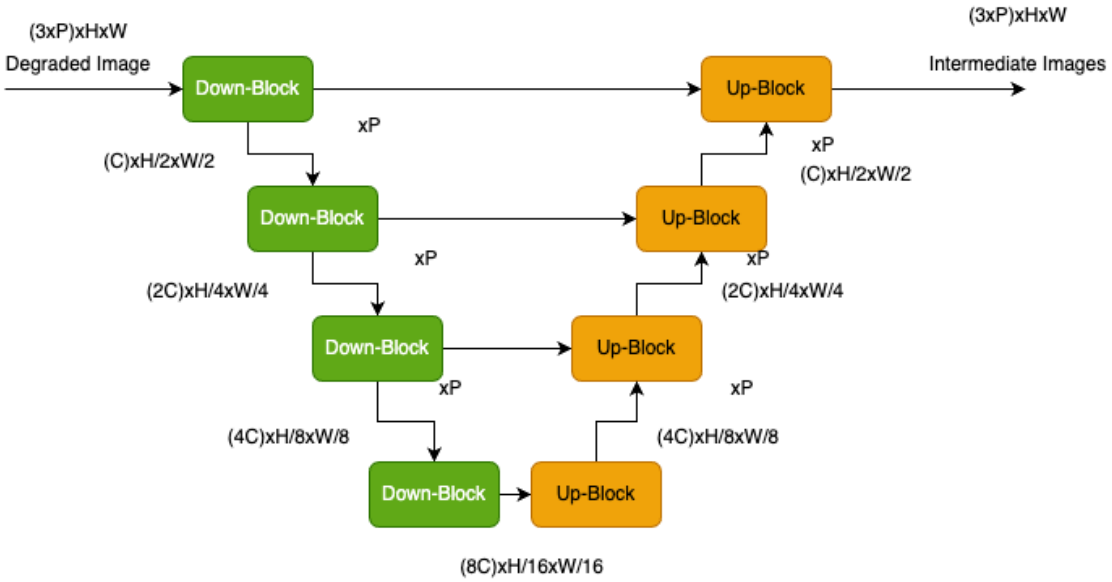

Figure 2.4: General U-Net Architecture. It consists of an encoding path shown here in green. Orange is the decoding path with skip connections. The last layer is the bottleneck.

U-Net architecture is a special class of neural network architecture that was originally proposed by Ronneberger, Fischer, and Brox (Ronneberger et al.) and was specifically designed for bio-medical image segmentation. It has a distinct U-shape structure with a contracting path, a bottleneck, and an expanding path. The contracting path downsamples the input image through a series of convolutional and pooling layers, progressively decreasing spatial resolution while extracting increasingly abstract features and capturing global context. Bottleneck acts as a bridge between the contracting and expansive paths, retaining crucial information from the downsampled representations. Expansive Path path utilizes transposed convolutions to upsample the feature maps, gradually restoring spatial resolution. Crucial to its success are the skip connections that directly link corresponding

levels of the contracting and expansive paths. These connections ensure the flow of high-resolution information from the contracting path to the expansive path, enabling precise localization and accurate segmentation of objects, even those with fine details.UNet excels at capturing both global and local features, leading to restored images with high fidelity and minimal artifacts.

2.5 Image Quality Metrics

Image quality metrics are quantitative measures designed to assess the fidelity and perceptual quality of images. These metrics play a crucial role in evaluating the performance of image processing algorithms, compression techniques, and other image-related applications. We are going to talk about some of the objective measures that allow us to evaluate quantitatively how well our network is performing compared to the rest.

2.5.1 Peak Signal to Noise Ratio

Peak Signal to Noise Ratio (PSNR), is a metric commonly used to evaluate the quality of an image or video compression algorithm. It measures the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. In image or video compression, PSNR is often used to quantify the reconstruction quality by comparing the original and compressed images. The PSNR is defined as below

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{2.5}$$

Here, MAX represents the maximum possible pixel value in the image (e.g., 255 for an 8-bit image), and MSE is the Mean Squared Error, which quantifies the average squared

difference between corresponding pixel values in the original and compressed images.

In essence, a higher PSNR value indicates a lower level of distortion or loss in the compressed image, implying a higher-quality reconstruction. The logarithmic nature of the scale emphasizes perceptually relevant differences, making PSNR a valuable tool for assessing the performance of compression algorithms in maintaining image fidelity.

2.5.2 Structural Similarity Index Measure

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.6}$$

where: (2.7)

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.8}$$

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2} \qquad \sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2} \tag{2.9}$$

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{2.10}$$

$$C_1 = 0.01 \qquad C_2 = 0.03 \tag{2.11}$$

SSIM stands for Structural Similarity Index Measure (SSIM), and it is a metric used to measure the similarity between two images. It was designed to assess the quality of images or videos by comparing their structural information, taking into account luminance, contrast, and structure.

The SSIM index produces a value between -1 and 1, where 1 indicates perfect similarity between the two images. Higher SSIM values correspond to more similar images, while lower values indicate greater dissimilarity. A value of 0 means no similarity.

The SSIM metric considers three components:

- Luminance (L): Represents the brightness of the image.
- Contrast (C): Measures the difference in intensity between pixels.
- Structure (S): Examines the spatial patterns and textures in the images.

The formula for SSIM is quite complex and involves comparing local patterns of pixel intensities. While SSIM is widely used, it's important to note that it may not always align with human perception of image quality, especially in certain cases where human judgment might differ from what SSIM indicates.

In image and video processing, SSIM is often used to evaluate the performance of compression algorithms, denoising techniques, and other image-processing applications. It provides a quantitative measure of how well the processed image retains the quality of the original.

2.5.3 Learned Perceptual Image Patch Similarity

Learned Perceptual Image Patch Similarity (LPIPS), is a powerful loss function employed in image restoration tasks. It goes beyond traditional pixel-level metrics, which focus solely on minimizing the difference between individual pixel values, by incorporating perceptual similarity based on human visual perception. This allows LPIPS to penalize image distortions that are visually noticeable but may not be readily apparent in pixel-level comparisons.

LPIPS utilizes a pre-trained deep convolutional neural network (CNN) to extract features from the image. The CNN is typically VGG16, but other models can be used. The

features are extracted at different scales, capturing both low-level and high-level information.

The image is divided into overlapping patches, and the features extracted for each patch are used to calculate a similarity score between the restored and reference images. The similarity score can be calculated using various metrics, such as cosine similarity or l2-norm.

It's crucial to note that, in the context of LPIPS, a lower score implies superior image quality. This metric serves as a valuable tool in image restoration tasks by incorporating human perceptual factors into the evaluation process, providing a more nuanced and accurate assessment of the visual quality of restored images.

Chapter 3

Shift Variant Image Blur Model

3.1 Introduction

Practical Imaging Systems such as lenses, telescopes, microscopes, etc tend to have imperfections that can be inherent or man-made. These imperfections cause deviations from the ideal optical behavior and cause a loss in image quality. These deviations are called Aberrations. There are types of aberrations that are space-variant, i.e., the PSF(Point Spread Function) or the impulse response is different at different points in the image.
The focus of our research has been to characterize and simulate this space-variant aberration and develop an effective image restoration model that takes into account the impulse response of the degraded image and efficiently deconvolves the degraded image to obtain the high-resolution image. In this chapter, we deal with the process of simulating space variant aberrations on images and compute it efficiently. We use Incoherent Image formulation using Fourier Transform to model our simple single lens system with a space-variant pupil function and implement a method called Overlap-and-Add to generate shift variant images efficiently.

3.2 Scalar Diffraction Theory and Beam Propagation

3.2.1 Rayleigh-Sommerfield diffraction integral

The propagation of light as an electromagnetic wave is governed by Maxwell's equation. Let us consider a 2D source plane or an illuminated aperture where the electric field distribution is defined and bounded. S is the area of the extended source. Now according to Huygens principle, each extended source can be seen as a collection of an infinite number of point sources which each generate spherical wavelets. The contributions of all wavelets are then summed at the coordinate x', y'. The "Rayleigh-Sommerfield diffraction integral" is expressed as

$$E(x', y', z) = \frac{z}{i\lambda} \iint_S E(x, y, 0) \frac{\exp\left[-ik\sqrt{((x-x')^2 + (y-y')^2 + z^2)}\right]}{((x-x')^2 + (y-y')^2 + z^2)} dxdy \tag{3.1}$$

The coordinates (x', y') represent the points in the observation plane and the (x, y) coordinates represent a point in the source plane. The integral is defined over the bounded region S. The source and the observation plane are parallel to each other so that z is the perpendicular distance between the two parallel planes.

3.2.2 Fraunhofer approximation

Calculating the Electric field using the above integral is computationally very expensive. Fortunately, we can use approximations to simplify the integral into more computationally feasible solutions. Let us consider the radial distance r of a point in the observation plane from the source plane. Using the binomial expansion we can simplify the argument in the

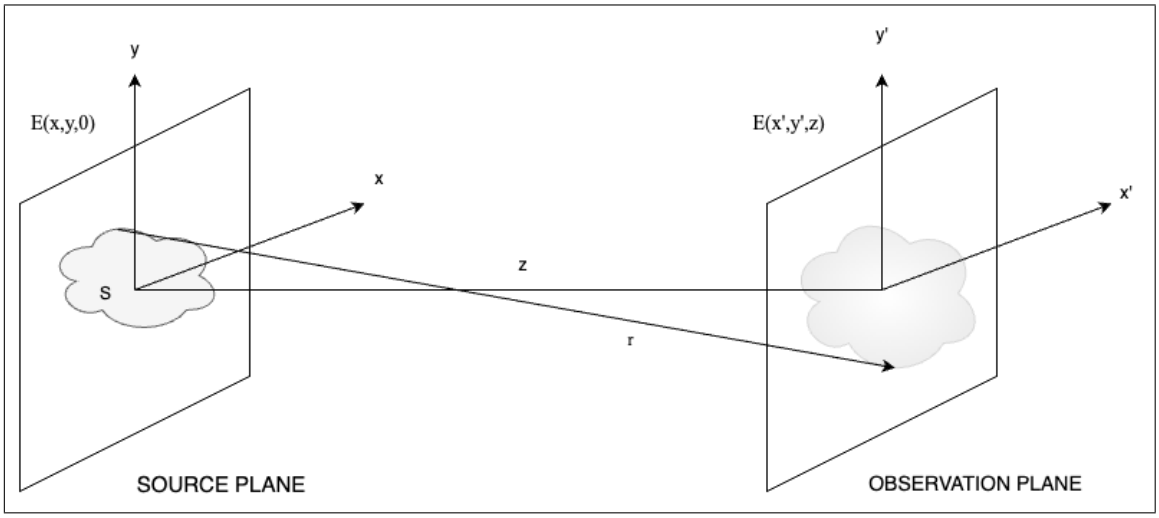

Figure 3.1: Beam Propagation

exponent:

$$k\left(z^2+(x-x')^2+(y-y')^2\right)^{1/2} \approx kz\left(1+\frac{(x-x')^2}{2z^2}+\frac{(y-y')^2}{2z^2}\right) \tag{3.2}$$

$$kz\left(1+\frac{(x-x')^2}{2z^2}+\frac{(y-y')^2}{2z^2}\right) = kz+\frac{k(x^2+y^2)}{2z}-\frac{kxx'}{z}+\frac{k(x'^2+y'^2)}{2z}-\frac{kyy'}{z} \tag{3.3}$$

In the denominator, the distance r between the two planes is usually dominated by z, so we can safely assume $r \approx z$. If we approximate the chirp term involving x and y terms inside the integral to unity, it implies that

$$z >> \left(\frac{k(x^2+y^2)}{2}\right)_{max} \tag{3.4}$$

Then our expression simplifies to that of a simple Fourier transform of the source field.

$$E(x', y', z) = \frac{1}{i\lambda z} e^{ikz} e^{\frac{i(x'^2+y'^2)}{2z}} \iint_S E(x, y, 0) e^{-ikxx'/z} e^{-ikyy'/z} dxdy \tag{3.5}$$

This regime is often called the "far field" and the integral is called the Fraunhofer diffraction integral.

3.2.3 Convolutional Approach to Image Formation

Using the Fraunhofer approximation and ensuring that we are in the far field regime we can safely use the Fourier Transform to get the Electric field of the source plane in the geometric image plane

$$E_i(x_i, y_i) = -\frac{1}{\lambda^2 do di} \int\int A(x_l, y_l) e^{-ik(x_i - Ma)x_l/di} e^{-ik(y_i - Mb)y_l/di} \, dx_l \, dy_l \tag{3.6}$$

$$= -\frac{1}{4\pi^2 di \, do} \int\int A \frac{k_{l,x} di}{k} \frac{k_{l,y} di}{k} e^{-ik_{l,x}(x_i - Ma)} e^{-ik_{l,y}(y_i - Mb)} \, dk_{l,x} \, dk_{l,y} \tag{3.7}$$

$$\equiv h(x_i - Ma, y_i - Mb) \tag{3.8}$$

The image field can then be expressed as a collection of point objects inside the bounded source plane, each contribution adding to an integral given below

$$E_i(x_i, y_i) = \iint E_o(x_o, y_o) \, h(x_i - Mx_o, y_i - My_o) \, dx_o \, dy_o \tag{3.9}$$

where M is the magnification factor and x_o and y_o are the source coordinates and x_i

and y_i are image plane coordinates. Let us make the following substitution $\tilde{x_o} = Mx_o$ and $\tilde{y_o} = My_o$ and $E_g(\tilde{x_o}, \tilde{y_o}) = E_o(x_o, y_o)/M^2$ then the above equation can be rewritten as

$$E_i(x_i, y_i) = \iint E_o(\tilde{x_o}, \tilde{y_o})\, h(x_i - \tilde{x_o}, y_i - \tilde{y_o})\, d\tilde{x_o}\, \tilde{y_o} \tag{3.10}$$

where is E_g is the geometric image is a magnified version of the object image where M is the scaling factor.

3.3 Incoherent Image Formation

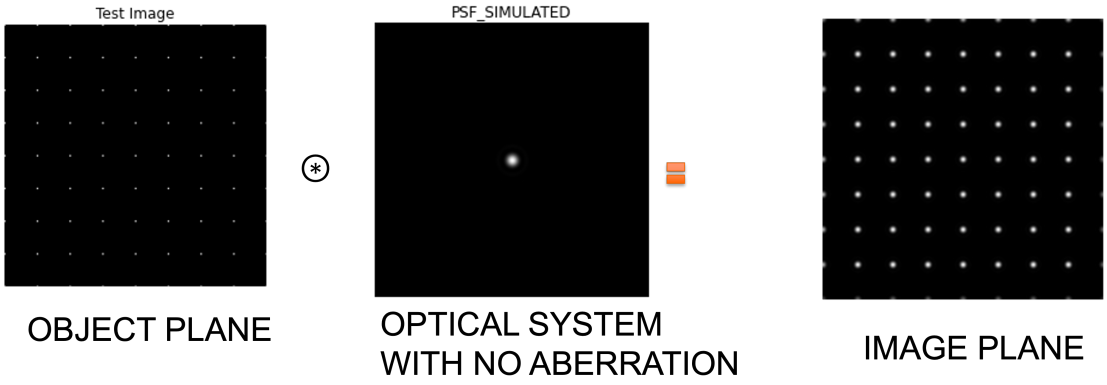

Figure 3.2: Incoherent Image Formation

The image usually measured by the sensor is the irradiance of the source plane and not its electric field. Irradiance is given by the time average of the value, which depends on the sensors' integration time. In incoherent illumination the phase of the field is random so there is no correlation between the different phases but with itself. We can use the above formulas to show that the image irradiance is dependent on the object irradiance and the incoherent impulse response which is also called the Point Spread Function (PSF) of the

imaging system by the formula

$$I_i(x_i, y_i) = \iint I_g(\xi, \eta)\, h(x_i - \xi, y_i - \eta)^2 \, d\xi \, d\eta \equiv I_g(x_i, y_i) * |h(x_i, y_i)|^2 \tag{3.11}$$

The Fourier transform of the PSF called the Optical Transfer Function or (OTF) converts the convolution integral for the shift-invariant case into a multiplication operation.

$$I_i(x_i, y_i) = \mathscr{F}^{-1}\left(\mathscr{F}(I_g(x_i, y_i) \times \mathscr{F}(|h(x_i, y_i)|^2)\right) \tag{3.12}$$

The result of such an operation is shown in 3.2.

3.4 Seidel Aberrations

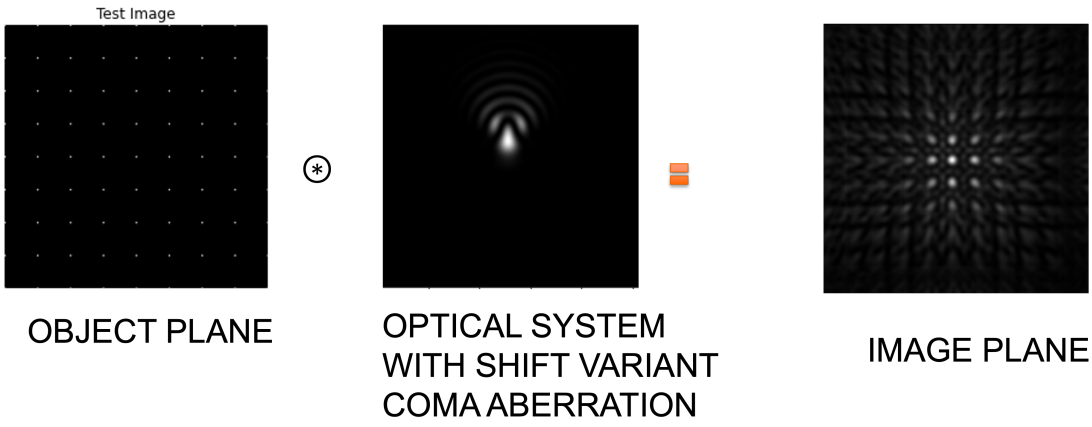

Figure 3.3: Effect of coma aberration on a star field image. Please note the difference in blurring and the orientation of the blur as it changes radially in strength. Output cannot be generated using a single convolution operation.

Seidel Aberrations, also known as primary or third-order aberrations, is a set of five fundamental optical aberrations that arise due to the imperfect focusing of light rays by an optical system. These aberrations were mathematically characterized by Ludwig von Seidel in the mid-19th century and are considered the most significant aberrations in optical systems.

These aberrations include spherical aberration, coma, astigmatism, field curvature, and distortion. Spherical aberration occurs when light rays passing through the outer portions of a lens focus at different points than those passing through the center, leading to blurred images. Coma 3.3results from off-axis light rays converging at different focal points, causing comet-like smearing of point sources away from the optical axis. Astigmatism occurs when light rays from a point source are not uniformly focused in two perpendicular meridians, resulting in distorted images. Field curvature causes curved focal planes, making it challenging to achieve sharp focus across the entire image. Distortion leads to a misrepresentation of object shapes, manifesting as barrel or pincushion distortions.

The causes of these aberrations are often linked to the inherent imperfections of optical surfaces, such as lens curvature and refractive index variations, which compromise the ideal convergence of light rays and contribute to image degradation. Addressing Seidel aberrations involves meticulous optical design and corrective measures to minimize these distortions and enhance the overall imaging performance of optical systems. We will show in later sections the nature of these aberrations and how simulation models can generate these aberrations on unblurred sharp images.

3.5 Shift Variant Blur Simulation

3.5.1 Introduction

Computational simulation of shift-variant blur is a difficult problem to solve. Simulating shift-variant blur involves using different blur kernels for different parts of an image. This requires more complex algorithms and computations compared to the simpler shift-invariant blur models. Modeling can introduce artifacts and edge effects, especially when transitioning between different blur regions. Managing these artifacts while maintaining image quality is a non-trivial task.

Shift-variant blur is not uniform across the image, making it challenging to define a single blur kernel. The characteristics of blur may vary based on factors like depth, motion, or other scene-specific properties. In our research, we use primary Seidel aberrations to generate shift-variant blur.

Multiple strategies have been proposed by researchers in computationally simulating shift-variant blur. Piecewise constant PSFs assume the PSF is shift-invariant for a particular partition of the image. Nagy and O'Leary (1998) proposes the opposite order of convolving then weighting different images. We follow the method described in Hirsch et al. (2010).

Our strategy to generate shift-variant blur is to use primary Seidel aberrations to vary the PSF model. The use of the Overlap-and-Add method as described in Hirsch et al. (2010). The next sections will expand on the algorithm followed.

3.5.2 Point Spread Function Model

Optical aberration is the distortion that is caused due to imperfections in the optical system for example an asymmetrical lens can cause the optical path length to increase or

decrease at some regions which would cause a deviation in the wavefront of the Gaussian spherical wave. We propose to simulate optical aberration by using a circular aperture with a wavefront aberration determined by the aberration polynomial with five primary terms also commonly called Seidel aberrations. The aberration polynomial is given by

$$W\left(h;\hat{x},\hat{y}\right)=W_{d}\left(\widehat{x^{2}}+\widehat{y^{2}}\right)+W_{040}\left(\widehat{x^{2}}+\widehat{y^{2}}\right)^{2} \tag{3.13}$$

$$+W_{131}h\left(\widehat{x^{2}}+\widehat{y^{2}}\right)\hat{x}+W_{222}h^{2}\widehat{x^{2}}+W_{220}h^{2}\left(\widehat{x^{2}}+\widehat{y^{2}}\right)+W_{331}h^{3}\hat{x}$$

where $\widehat{x}$ and $\widehat{y}$are the exit pupil coordinates and h is the object height.The pupil plane is normalized to a maximum value of 1. The image plane is normalized to 1 such that h represents the factional height of a single-point object. For any point in the object plane, we need to do a rotational transform to align the pupil coordinate system to the object plane center. In other words, for each point in the object plane, there will be x and y components to the aberration which can be found if we did the following coordinate transform.

$$X_{r}=\hat{x}cos\left(\beta\right)+\hat{y}sin\left(\beta\right),\ Y_{r}=-\hat{x}sin\left(\beta\right)+\hat{y}cos\left(\beta\right),\ \beta=arctan\hat{y}/\hat{x} \tag{3.14}$$

The Point Spread function can then be found out by multiplying the aperture function $A(\widehat{x},\widehat{y}$ with the phase of the function. Assuming all the other constants to be equal to unity we can formulate the equation as:

$$k(\widehat{x},\widehat{y},u,v)=|\mathscr{F}(A(\widehat{x},\widehat{y})e^{i\kappa W(u,v)})|^{2} \tag{3.15}$$

where $\mathscr{F}$ is the Fast Fourier transform of the image. As an example, we can display the

effect of coma-aberration by just considering the effect of the third coefficient in the wave aberration polynomial and setting others to zero as in 3.3.

3.5.3 Image Blur Model

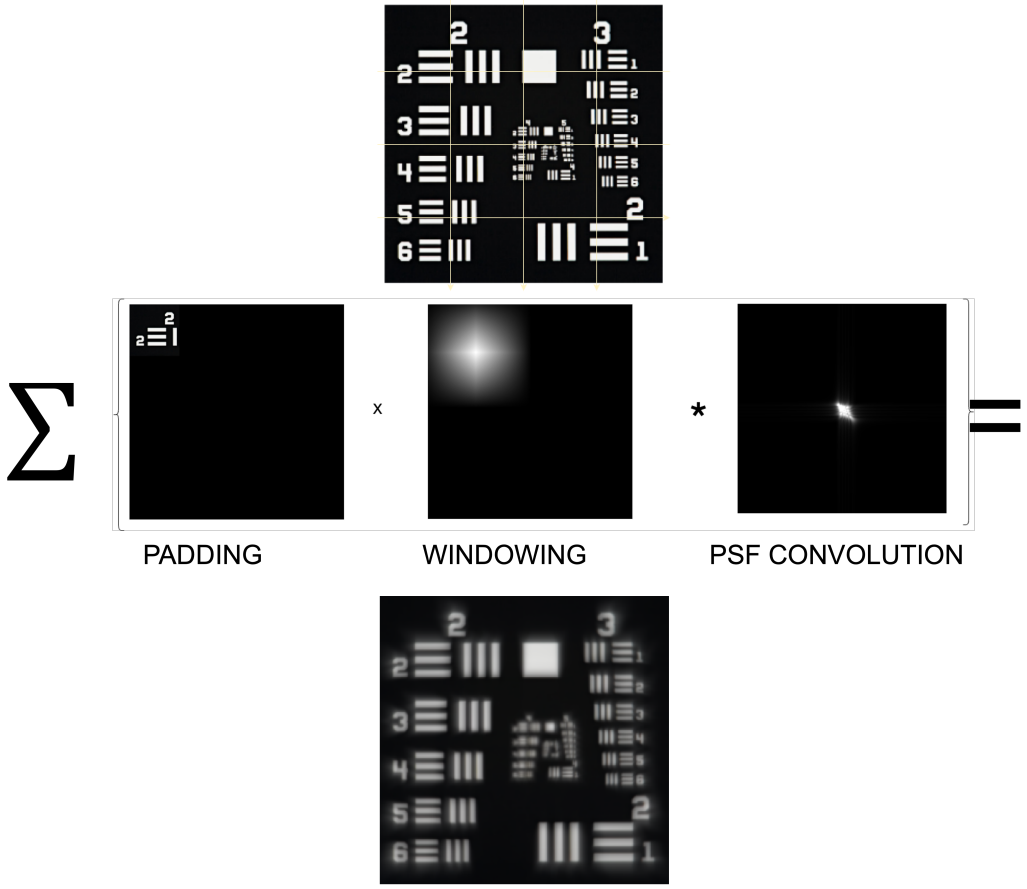

Figure 3.4: Overlap and Add Method: The image is divided into overlapping patches with a step window function which is zero everywhere else except the patch and multiplied with a Bartlett window function that is convolved with the PSF filter for the particular patch. Then all the patches are summed up.

Since simulating space-variant image blurring is computationally expensive, we make use of the method proposed in Hirsch et al. (2010) called the overlap-add method(OLA). We divide the image into overlapping patches where we assume the PSF to be shift-invariant and use a Bartlett window function to dampen the borders of each patch to suppress border artifacts. The amount of overlap is kept at 0.5 which implies the step size of a window function going over each patch has to be half of the PSF size. For our purpose we choose a window size of 64 pixels.

Then the shift variant operation turns into a sum of shift invariant operations where for each patch the PSF is determined by the central coordinate of the patch.

Mathematically, let p be the number of overlapping input image patches f of length m. We define the window functions as $w^{(i)}$ and a PSF function $k^(i)$ of length l. We can then formulate it as follows:

$$g(x) = \sum_{i=0}^{p-1} \sum_{j=0}^{l-1} k_j^i w_{x-j}^i f_{x-j} \quad for\ 0 \leq x < m \tag{3.16}$$

The key requirement is that the sum of window functions for all pixels within a patch must equal one. The reason we chose the Bartlett window is the fact that it is a triangular function that satisfies the summation property. This condition is vital to prevent artifacts from emerging in the overlapping regions of adjacent patches.

$$\sum_{r=0}^{p-1} w_i^r = 1 \quad \text{for} \quad 0 \leq i < m \tag{3.17}$$

3.6 Experimental Results

The simulation is done using a circular aperture with wavelength value kept constant for all three channels at $0.55\mu m$.We select an exit pupil distance of 100mm and an image plane length of 1mm to exaggerate the amount of blur generated in the image. Since normal values wouldn't generate as much shift variance in the blur and qualitatively it will be hard to distinguish, we adopt such a practice. The sample size is equal to the PSF size which is 256 pixels, giving us a sampling interval of $\frac{10^{-}3}{256}$ m/sample.

3.6.1 Experiment-1

We are going to isolate all the primary aberrations so that we can see the effects of them on sharp images separately. We use 5 wavelengths of aberration coefficient for the aberration coefficient that we want to visualize and force all the other aberration coefficients to 0. The defocus and spherical aberration in 3.5 are shift-invariant and can be modeled using a single PSF. The rest of the degradation is different at different regions of the network as shown in figure 3.6.

Table 3.1: Seidel Coefficient Values for f/5 lens(from ZEMAX)

Coefficient	**Value**[a]
W_d	0
W_{040}	4.963λ
W_{131}	2.637λ
W_{222}	9.025λ
W_{220}	7.536λ
W_{311}	0.157λ

[a] $\lambda = 0.55\mu m$.

3.6.2 Experiment-2

Figure 3.9: Blurred images generated using Zemax coefficients for f/5 lens

We use the Seidel polynomial coefficient provided by ZEMAX for an f/5 lens as given in Chapter 8 of Voelz (2011). We use the same setup as before but instead, the aberration coefficients are non-zero values and the total effect is mixed. The resultant PSF is shown in 3.7. The PSF has a significant amount of distortion from the ideal airy-ring pattern because of Seidel aberrations. The wavefront of the corresponding region 3.8 also shows the deviation from the ideal spherical wavefront is also shown. These will be standard values that we are going to use to generate our dataset.

3.7 Discussion

Our shift variant simulation model uses Seidel aberration to generate shift-variant blur. Then we use the overlap-and-add method to generate shift-variant efficiently. However, there are limitations to the simulation. The quantization of PSFs at central coordinates of patches leads to border artifacts. The realistic shift variant models would become more complicated as we computationally encapsulate depth effects. Achieving a high level of realism in simulating shift-variant blur often involves complex models and algorithms. However, there is a trade-off between realism and computational efficiency, and finding a balance is crucial for practical applications. We have tried to minimize border artifacts but we need to be cautious of the fact that two much of aberration might break the shift-variant model as the difference in PSfs between two regions becomes significant enough to cause border artifacts to become qualitatively visible.

Figure 3.5: Shift Invariant aberrations in Seidel Polynomial. The PSF model is same everywhere and can be represented by a single convolution

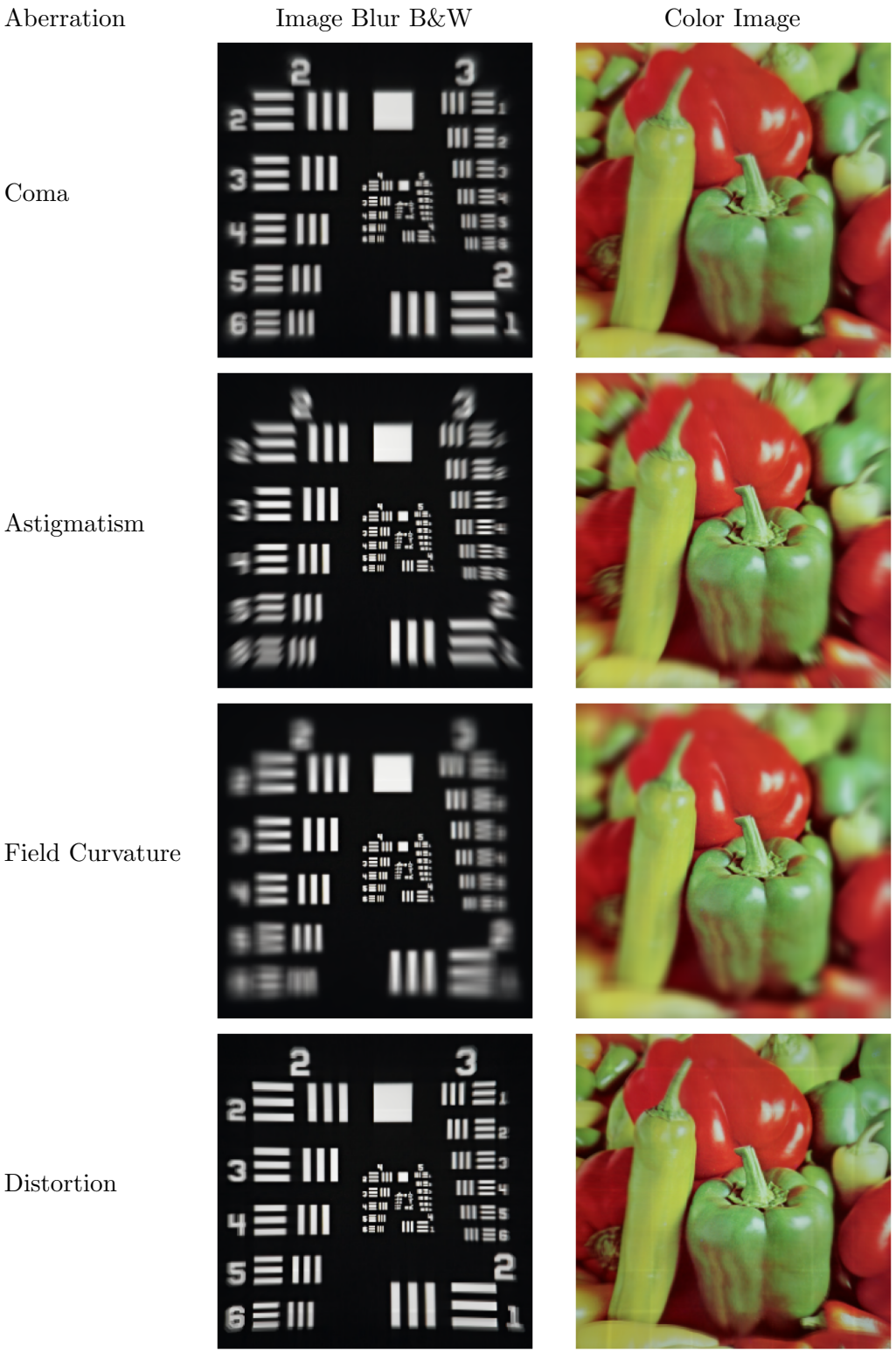

Figure 3.6: Shift Variant Aberrations. All primary Seidel Aberrations that have image plane dependency have variable blurring effects across the images.

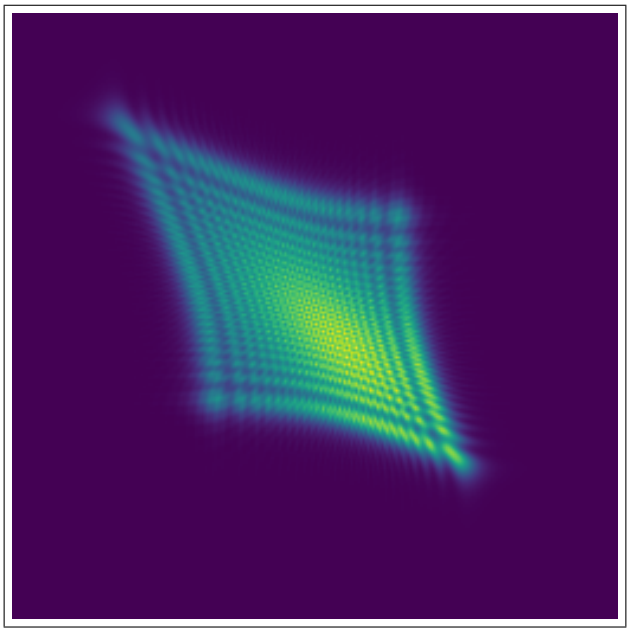

Figure 3.7: PSF at the normalized image coordinates u=1 v= 1 for f/5 lens and with aberration coefficients as discussed

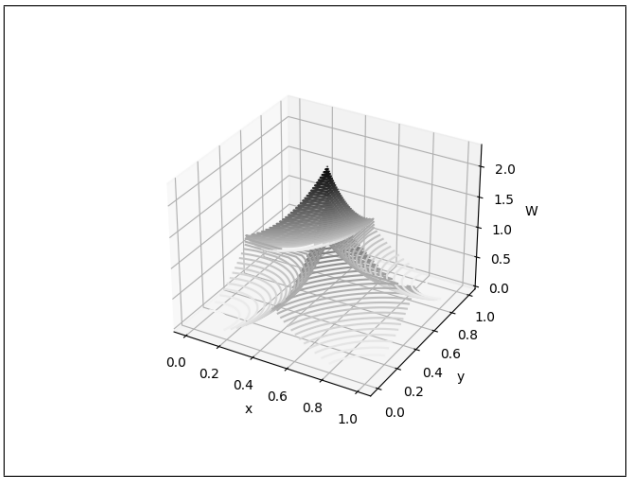

Figure 3.8: Wavefront of f/5 lens provided by ZEMAX

Chapter 4

Research

4.1 Introduction

Deep Learning methods have demonstrated remarkable potential in the domain of low-level image processing, offering advantages over conventional image deconvolution algorithms, especially in handling complex scenarios and incorporating intricate noise models. In Chapter 2, we discussed two overarching methods: Non-Blind deconvolution and Blind Deconvolution. For our specific dataset, where the Point Spread Function (PSF) is known, we aim to leverage this knowledge to enhance the effectiveness of our image deconvolution through deep learning models.

In this chapter, we first look into research that deals with shift-variant degradation in both blind and non-blind manner. ShiVaNet addresses shift variance by employing multiple Wiener filters for non-blind deconvolution. The model capitalizes on the spatially varying characteristics of degradation present in the dataset. Notably, the learnable Wiener Deconvolution module utilizes Simplified Channel Attention (SCA) and Transposed Attention to enhance intermediate features, ensuring effective information propagation.

Furthermore, we enhance our U-Net refinement module by integrating ConvNext-V2, a novel architecture that improves feature diversity through the use of Global Response Normalization. This augmentation aims to capture a broader range of image features, contributing to the overall effectiveness of our deconvolution process.

The outcomes of our experiments with ShiVaNet are thoroughly discussed in the Results section, shedding light on the model's performance in comparison to existing methodologies. Additionally, we analyze the impact of each architectural component on the overall effectiveness of shift-variant image deconvolution.

Our main contributions are summarized below

- **Multi-Wiener Transposed Attention**: We use Simpified Channel Attention (SCA) and Transposed Channel Attention on the intermediate images generated after the learnable Wiener Deconvolution module to improve global context feature mapping.
- We propose to merge pixel-wise linearly using a depthwise convolution before sending to U-Net refinement.
- **U-Net Fusion Block**: We propose a fusion block that incorporates ConvNext-v2 in conjunction with Transposed Channel Attention with skip connections and a parallel connection of the ConvNext-v2 module with SCA. The lateral inhibition of features by normalization caused by ConvNext-v2 improves feature diversity.

4.2 Related Work

Image Deconvolution research has primarily focused on Shift Invariant Deconvolution or estimating the blurring function using deep neural networks. Traditional image processing algorithms that deal with shift variant deconvolution usually use a coordinate transforma-

tion to convert the shift variant problem into a shift-invariant one. In Robbins and Huang (1972), the authors use Mellin Transform to solve shift-variant coma aberrated systems and then extend it to a wider class of similar shift-variant systems. The authors decompose the problem of inverse filtering into a distortion or a coordinate transformation of the image plane. The inverse filtering problem transforms from a shift-variant to a shift-invariant one. After the shift-invariant deconvolution, the image is again distorted at the output end to recover the original sharp image. Estatico and Di Benedetto (Estatico and Di Benedetto) also follows a similar method and proposes an algorithm to find a coordinate transform that converts the structured shift-variant PSF into a shift-invariant one.

Modern deblurring methods leveraging deep learning often adopt a generalized image restoration approach. These methods go beyond traditional algorithms by utilizing deep neural networks to establish intricate relationships between pixels and their neighborhoods. The primary objective is to determine an inverse solution to the image restoration problem based on these learned relationships.KBNet Zhang, Li, Shi, He, Song, Wang, Qin, and Li (Zhang et al.)is a novel deep-learning architecture designed for image restoration tasks. It leverages a kernel basis attention (KBA) module to efficiently capture long-range dependencies and spatial-invariant features. The KBA module forms the core of KBNet and learns a set of basis kernels that can be adaptively combined to represent the global context of an image. Additionally, a multi-axis feature fusion (MFF) block is employed to encode and fuse channel-wise, spatial-invariant, and pixel-adaptive features, providing a richer representation for image restoration. This enables KBNet to achieve state-of-the-art performance on various image restoration benchmarks, including denoising, deraining, and deblurring while requiring less computational cost than previous methods. Since it is a blind deconvolution method, it doesn't require PSF information. Similarly, NAFNet

Chen, Chu, Zhang, and Sun (Chen et al.) proposed a unique nonlinear activation-free network and replaced its multiplication within the channels of the same feature tensor to generate non-linearity in the network. It achieves 33.69dB PSNR on the GoPro dataset in deblurring problem.Restormer Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.) uses efficient Multi Dconv-head transposed channel attention to capture long-range pixel interactions and a gated Dconv feed-forward network to suppress less informative features. Image deblurring can also be seen as an image translation problem. Image translation is seen as finding a mapping function from an input distribution, in our case the blurred Images to unblurred sharp images.

Some models use generative adversarial networks to find image correspondences between blurred and unblurred images.Pix2Pix Isola, Zhu, Zhou, and Efros (Isola et al.) utilizes a conditional generative adversarial network (GAN) architecture, trained with paired data, to map input blurred images to desired unblurred images. This makes it suitable for tasks like image colorization, object removal, and style transfer. CycleGAN Zhu, Park, Isola, and Efros (Zhu et al.), on the other hand, employs two unpaired GANs trained cyclically, enabling translation between unpaired domains without requiring explicit paired examples. Both can be utilized for deblurring applications.

The most relevant research that has been done on shift variant deconvolution has been Multi-Wiener Net Yanny, Monakhova, Shuai, and Waller (Yanny et al.). Building upon the work by Dong et al. (2021), the neural network model is a two-stage architecture comprising of a Multiple Wiener Deconvolution module and U-Net refinement step. The learnable Multiple Wiener Deconvolution uses learnable PSF's initiated by the optical system information to generate Wiener Deconvolved multiple intermediate images which are then combined and refined by the U-Net refinement step. We base our skeleton architecture on

this network.

4.3 Method

In this section, we define and highlight the important design modification made to the network. The input image size is kept fixed at 256x256 pixels and all computations and feature sizes are calculated based on that input image size.The skeleton architecture is inspired by Multi-Wiener Net as proposed by Yanny, Monakhova, Shuai, and Waller (Yanny et al.). We have a two stage model; the first stage is termed as Multi-Wiener Transpose Attention Block and the next stage is the U-Net Refinement Step. We introduce architectural change in both of the modules.The following section will discuss the modules that were incorporated into our architecture.

4.3.1 Multi-Wiener Transposed Attention Block

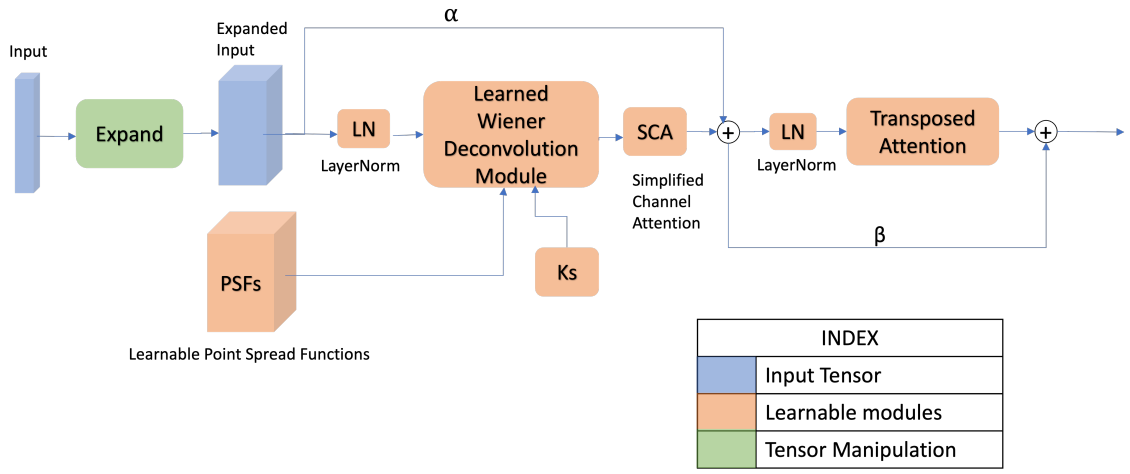

Figure 4.1: Architecture of Multi-Wiener Transposed Attention Block

The conventional closed-form approach of the Wiener-deconvolution algorithm can be made learnable by initializing them with the known model parameters and then learning using deep neural networks.Yanny, Monakhova, Shuai, and Waller (Yanny et al.) learns multiple Wiener filters and then couples it with a U-Net refinement step. We intend to use a similar configuration and learn multiple Wiener filters for our U-NET model in a configuration similar to Zhang, Li, Shi, He, Song, Wang, Qin, and Li (Zhang et al.).

Let $\mathbf{S}_i(\mathbf{u}, \mathbf{v})$ denote the Fourier transform of the input image $\tilde{X}_i(x, y)$, and $\mathbf{H}_i(\mathbf{u}, \mathbf{v})$ represent the Fourier transform of the PSF (Point Spread Function) $h_i(x, y)$. The Wiener deconvolution is then applied as follows:

$$\mathbf{S}_i(\mathbf{u}, \mathbf{v}) = \mathscr{F}\{\tilde{X}_i(x, y)\}, i = 1, \ldots, N \tag{4.1}$$

$$\mathbf{H}_i(\mathbf{u}, \mathbf{v}) = \mathscr{F}\{h_i(x, y)\}, i = 1, \ldots, N \tag{4.2}$$

$$\hat{\mathbf{P}}_i(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{H}^*{}_i(\mathbf{u}, \mathbf{v})\mathbf{S}_i(\mathbf{u}, \mathbf{v})}{|\mathbf{H}_i(\mathbf{u}, \mathbf{v})|^2 + \lambda_i} \tag{4.3}$$

Here, $\hat{\mathbf{P}}_i(\mathbf{u}, \mathbf{v})$ represents the estimated Fourier transform of the restored image, and λ_i is a regularization parameter associated with the i-th Wiener deconvolution operation.

We expand the input image into N copies equal to the number of PSF filters we want to learn and set up N PSF filter and regularization parameters for each of the Wiener deconvolution operations. The inverse Fourier transform of $\hat{\mathbf{P}}_i(\mathbf{u}, \mathbf{v})$ is denoted as $\hat{\mathbf{Q}}_i(x, y)$, and it is obtained as:

$$\hat{\mathbf{Q}}_i(x, y) = \mathscr{F}^{-1}\{\hat{\mathbf{P}}_i(\mathbf{u}, \mathbf{v})\} \tag{4.4}$$

We use layer normalization before passing it into a learnable Wiener Deconvolution module which is then coupled with Simplified Channel Attention. We add skip connections

to address the vanishing gradient issue and then couple it with the Multi-Dconv head Transpose Attention Block proposed in Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.). The following architectural modifications as shown in 4.1 are described below.

4.3.1.1 Simplified Channel Attention

The Simplified Channel Attention (SCA) module, proposed by Chen, Chu, Zhang, and Sun (Chen et al.), addresses the Wiener deconvolution information present in each channel of an intermediate image tensor. Motivated by the need for efficient interaction among channels to share information about aggregated Wiener deconvolution with variable PSF, the SCA module provides an effective mechanism for this channel-wise communication.

Mathematically, the SCA operation on an input tensor $\hat{\mathbf{Q}}_i$ is represented as:

$$\hat{\mathbf{V}}_i = SCA\{\hat{\mathbf{Q}}_i\} = \mathbf{Q}_i * \mathbf{W}pool\{\mathbf{Q}_i\}$$

Here, $\hat{\mathbf{V}}_i$ is the output of the SCA module applied to the i-th channel. The operation involves convolution ($*$) of the input tensor $\mathbf{Q}_i$ with a weight tensor $\mathbf{W}$, which encapsulates the pooling of information related to Wiener deconvolution with variable PSF. The Simplified Channel Attention module facilitates an efficient way for channels to collaborate and share pertinent information, enhancing the tensor's representation of aggregated Wiener deconvolution details.

4.3.1.2 Transposed Channel Attention

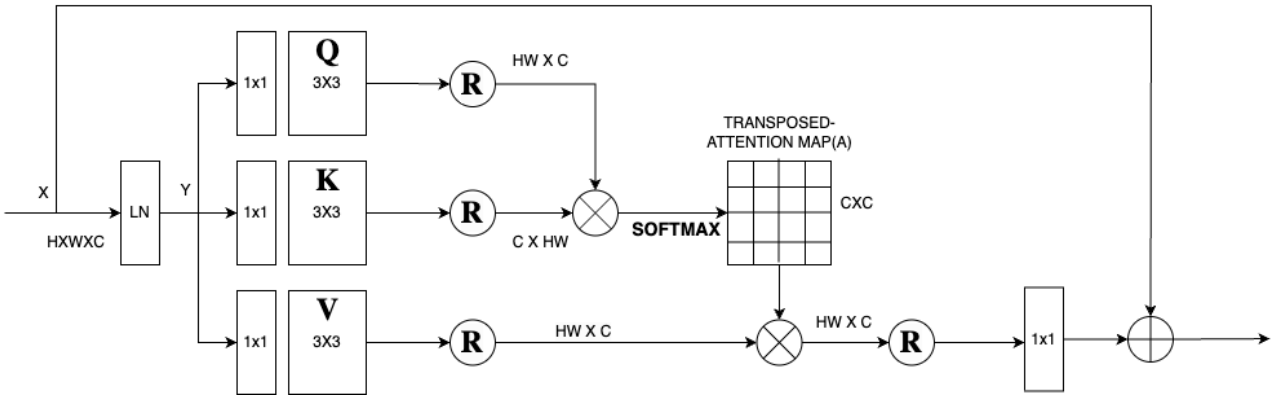

Figure 4.2: Architecture of Transposed Channel Attention Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.)

The Wiener deconvolution with a learned PSF only deconvolves perfectly for a local patch instead of globally for the whole image, we need the global context of the image in which self-attention is useful. We use the Transposed Attention Block because it is much more efficient to use self-attention across channels rather than spatially and it captures local context before computing the global attention map. The method is the same as described in Restormer paper Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.).

$$\hat{\mathbf{W}}_i = SCA\{\hat{\mathbf{Q}}_i\} + \alpha\mathcal{X} \tag{4.5}$$

$$\hat{\mathbf{Y}}_i = TransAttention\{\mathcal{LN}\{\hat{\mathbf{W}}_i\}\} + \beta\hat{\mathbf{W}}_i \tag{4.6}$$

4.3.2 U-Net Refinement Step

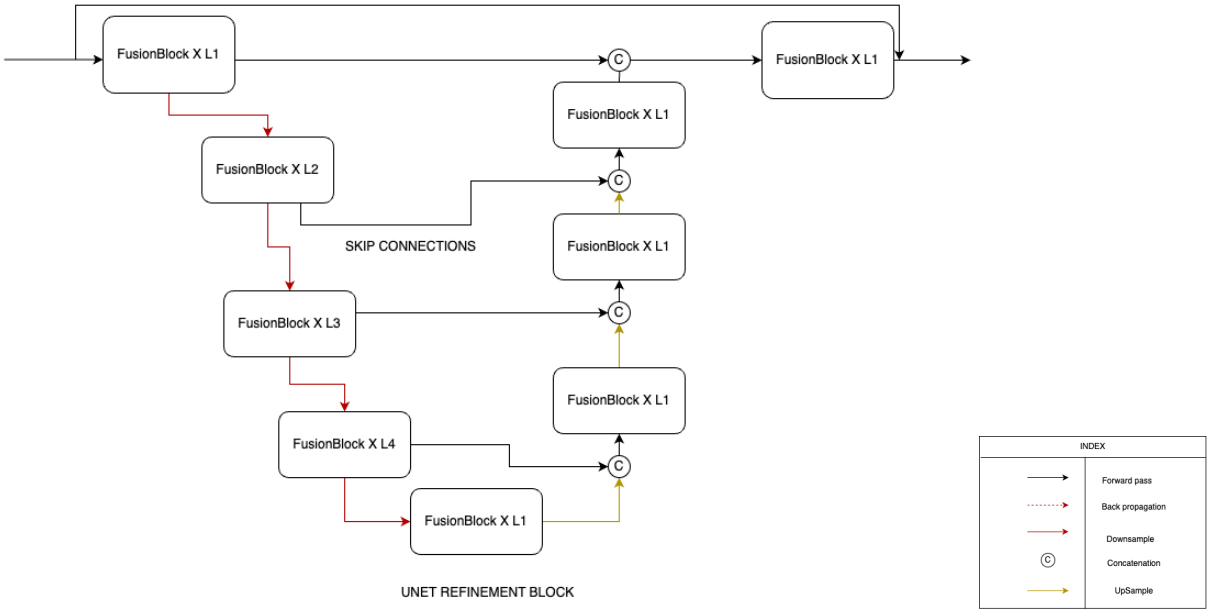

Figure 4.3: Architecture of U-Net Refinement Block

The first stage of our network generates intermediate images which are then combined using a point-wise convolutional layer. Then the image is fed to our refinement module shaped like a U-Net network. The fusion block of our U-Net architecture is used to process the input across different resolutions through down-sampling and up-sampling. The feature map resolution is increased similarly as Zhang et al. (2023), by the Pixel-Shuffle operation. Network Architecture details are given in figure 4.3.

4.3.2.1 U-Net Fusion Block

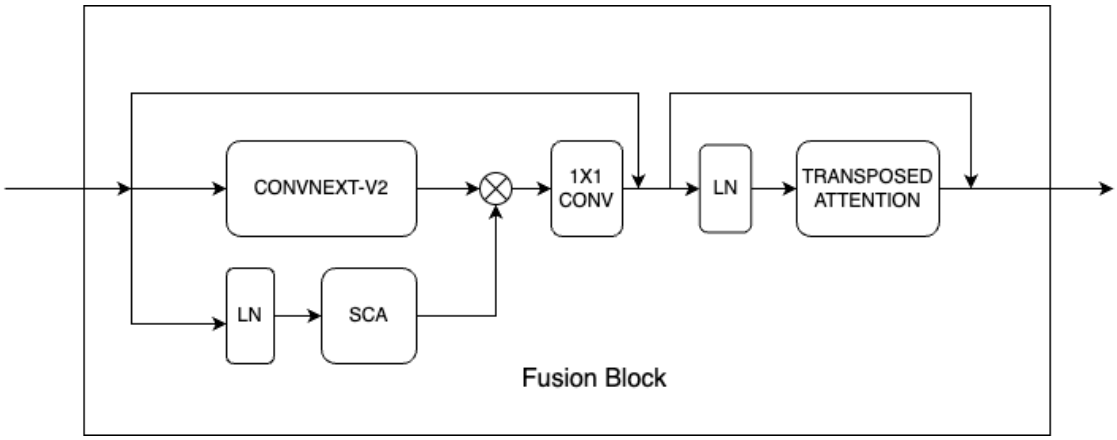

Figure 4.4: Architecture of Fusion Block. The Block is used in both the Encoder as well as Decoder Steps.

Our UNet fusion blocks are composed of ConvNext-V2 Woo, Debnath, Hu, Chen, Liu, Kweon, and Xie (Woo et al.), Simplified Channel Attention Module Chen, Chu, Zhang, and Sun (Chen et al.) and Transposed Channel AttentionZamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.). We use the ConvNext-V2 block as our primary block to capture spatially-invariant features. It utilizes Global Response Normalization(GRN) which increases the diversity of features extracted at each stage in our refinement block. It promotes feature competition across the feature channel. ConvNext-V2 also learns the relative position bias and allows us to inject spatial awareness into our network.

We also utilize the simplified channel attention to modulate our features across the channels. Both of the branches are then multiplied point-wise directly and then passed to a projection layer comprising of point-wise convolution layer. The point-wise multiplication acts as the non-linear activation in our network. Skip connections are provided to improve gradient flow. The final stage of our fusion block is a Multi-Dconv head Transposed At-

tention Block Zamir, Arora, Khan, Hayat, Khan, and Yang (Zamir et al.). The transposed attention map encodes the cross-covariance across the feature channels. We learn parallel separate attention maps.

The ConvNext-V2 block with Global Response Normalization (GRN) in conjunction with Transposed channel attention possibly forces the channels to increase cross-variance among the channel-wise features and learn global dependencies.

4.3.3 Stage-II Model Architecture

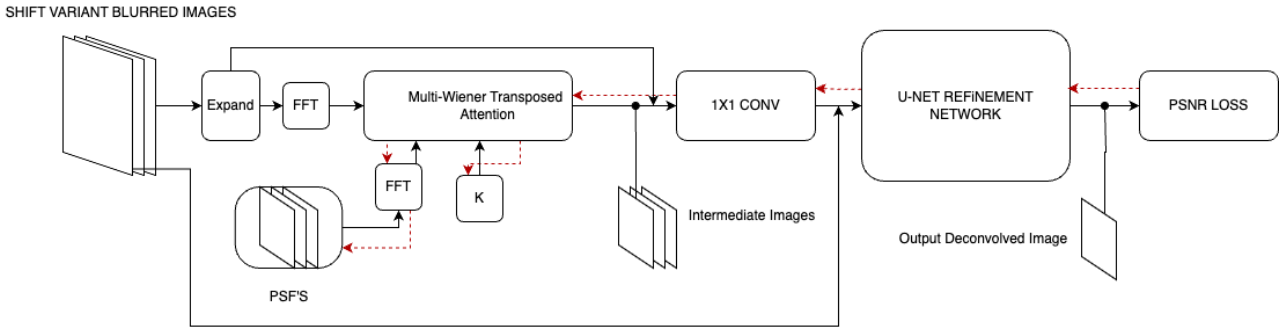

Figure 4.5: Proposed Architecture of ShiVaNet.Two-stage architecture composed of Multi-Wiener Transposed Attention and a U-Net Refinement Network coupled with a Depth-wise Convolution Layer. The red arrow indicates error back-propagation.

We adopted the two-stage architecture so that the first stage which incorporates PSF information can be used in the second stage to boost performance. Since blind image restoration is a well-studied problem, we can couple our proposed Multi-Wiener Attention module with a U-Net refinement network.

The first stage, Muti Wiener Transposed Attention (MWTA) incorporates the PSF information when the PSFs are initialized with the simulation coefficient values. When the network is allowed to learn the PSFs and regularisation term Ks are also allowed to learn

and converge to a better PSF and regularization term.

The intermediate images are combined with a depth-wise convolution layer with 1×1 kernel that condenses the multiple images into a linear combination of pixel intensities at each location.

The last stage is the U-Net refinement network that finds out patterns in the images to generate finer details in the images and reduce noise.

4.4 Experiments

The dataset generation, training, and implementation details are described in this section. We evaluate ShiVaNet on popular blind image restoration models and a non-blind one Multi-Wiener Net Yanny, Monakhova, Shuai, and Waller (Yanny et al.).

4.4.1 Dataset

We take the Go-Pro dataset and center-crop all the images to a square of 256 pixels. We add random rotations and random flips to the images. We also add Gaussian Noise to the images with a zero mean and a $\sigma = 0.0001$. The GoPro dataset is sampled and the forward blur model is applied to generate the shift-variant blur on the ground-truth. We generate 5000 training image pairs and 1000 test image pairs. We use the primary Seidel coefficient values calculated by ZEMAX for the f/5 lens. They are given in Table 3.1.

4.4.2 Training Details

We initialize our Learnable PSF tensor with the calculated ones from the forward blur model. Our training set is composed of 5,000 images taken from GoPro and blurred using our forward model. Image size is kept at 256×256, batch size is kept at 8. We use

16 16-channel PSF tensor to approximate the Shift Variant blur model. All models for comparative study were trained for 200 epochs. Patch sizes for the forward blur model are kept at 64 with a step size of 32. UNet refinement model encoder block numbers are given by $\{4, 4, 4, 8\}$ and decoder block numbers are given by $\{8, 4, 4, 4\}$ with 1 middle block. All other parameters are the same as Chen, Chu, Zhang, and Sun (Chen et al.).

We use the PSNR loss to train our model. The PSNR loss is given by:

$$\text{PSNR\_loss} = -10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right)$$

where MAX is the maximum possible pixel value and MSE is the Mean Squared Error.

4.5 Observation

Table 4.1: Shift Variant Deblurring comparisons on Test Dataset(using ZEMAX coefficients. 1000 images from GoPro Dataset)

Method Type	**Metric**		
	***PSNR*↑**	***SSIM*↑**	***LPIPS*↓**
Input	19.0732	0.4639	0.5181
KBNET	27.4901	0.8146	0.0958
NAFNET	27.4979	0.8118	**0.0204**
Multi-Wiener Net	19.5185	0.5696	0.1871
PIX2PIX	22.1650	0.6131	0.2092
CYCLE-GAN	23.6972	0.6934	0.1402
ShiVaNet(Ours)	**27.6808**	**0.8182**	0.0256

We compute PSNR(Peak Signal to Noise Ratio), SSIM(Structural Similarity Index), and LPIPS(Learned Perceptual Image Patch Similarity) scores for popular methods available in the literature. Table 4.1 shows that ShiVaNet outperforms the other methods

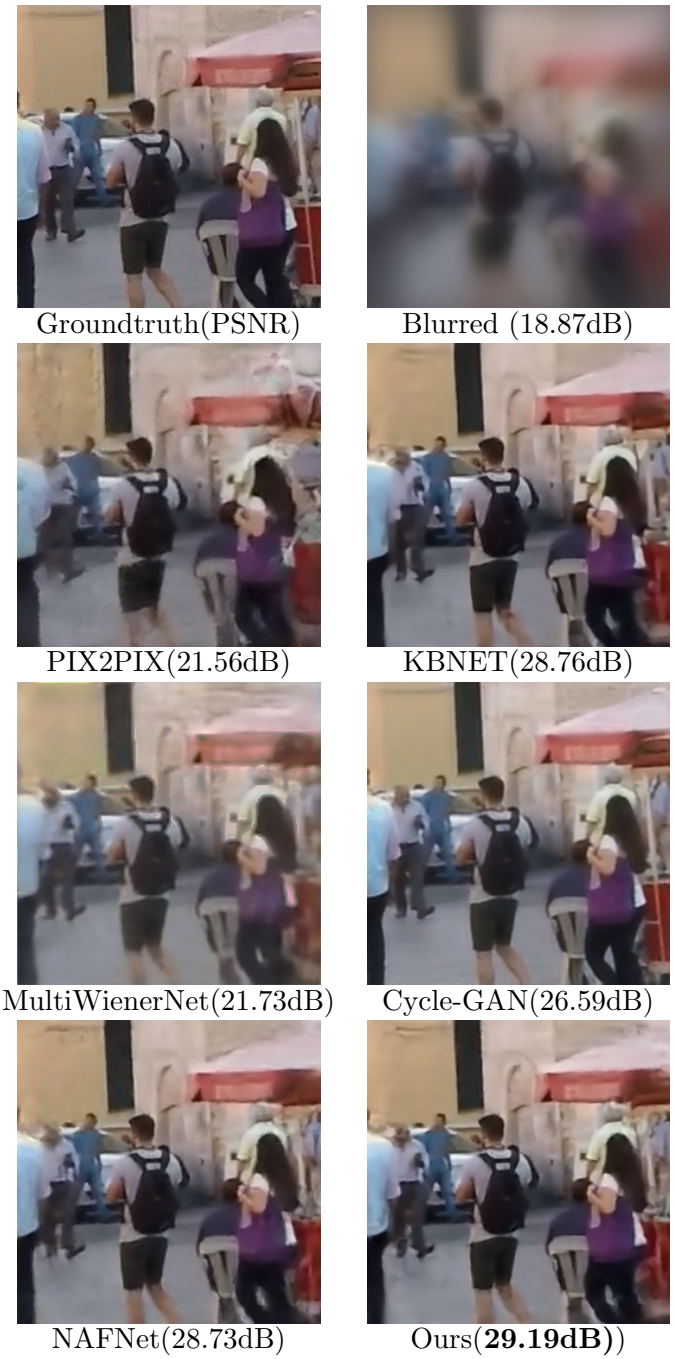

Figure 4.6: Sample Results from the GoPro dataset blurred using ZEMAX f/5 lens and individual PSNR value calculated for given method

concerning PSNR score and outperforms in SSIM score as well. Compared to NAFNet Chen, Chu, Zhang, and Sun (Chen et al.) our model provides a substantial gain of 0.2dB if the PSNR scores are compared.

4.6 Conclusion

In our research, we propose a novel method for restoring images degraded by shift-variant aberration. We have introduced design modifications to the Wiener-Deconvolution Module by incorporating Simplified Channel Attention and Transposed Channel Attention between the intermediate images. The intermediate images are then merged using a depthwise convolutional layer, and the resulting output is fed into a UNet Refinement Block that further enhances the signal-to-noise ratio of the image.

Our UNet refinement block includes a ConvNext-v2 block utilizing Global Response Normalization to increase feature diversity, along with a Transposed Channel Attention that extracts the global context of these features.

Our model outperforms(0.2dB improvement over Chen, Chu, Zhang, and Sun (Chen et al.)) best networks on blind image restoration in terms of metrics as well as image quality. It significantly outperforms Yanny, Monakhova, Shuai, and Waller (Yanny et al.) which is the backbone of our network.

Chapter 5

Future Directions

ShiVaNet demonstrates a noticeable improvement over the state-of-the-art in handling shift-variant aberrations, particularly when there is prior knowledge of the blurring function or Point Spread Function (PSF). An intriguing avenue for extending our model lies in addressing motion and rotational motion blur. In instances where the camera rotates around the principal optical axis, the resulting blur on the model exhibits a shift-variant nature. The blurring paths are shorter around the axis and elongate towards the edges of the image. We aim to explore the effectiveness of our Multi-Wiener Transposed Attention module in estimating the blurring operator, particularly when initialized with randomly generated PSF tensors.

Our dataset was generated with a fixed blur model, incorporating fixed Seidel aberration coefficients. To further enhance the model's adaptability, we propose training it on datasets with randomized Seidel aberration coefficients. This approach introduces variability in the intensity of shift-variant blur, and we hypothesize that the estimated PSF tensors will inherently encode valuable information about the dataset.

While simulating shift-variant blur computationally proves to be more efficient, it is

crucial to emphasize that our model should also be adept at processing blurs derived from real-world optical systems. Consequently, we plan to train our network using a realistic blur dataset to evaluate its performance and determine if it surpasses the current state-of-the-art in handling such complexities.

Appendix A

Appendix

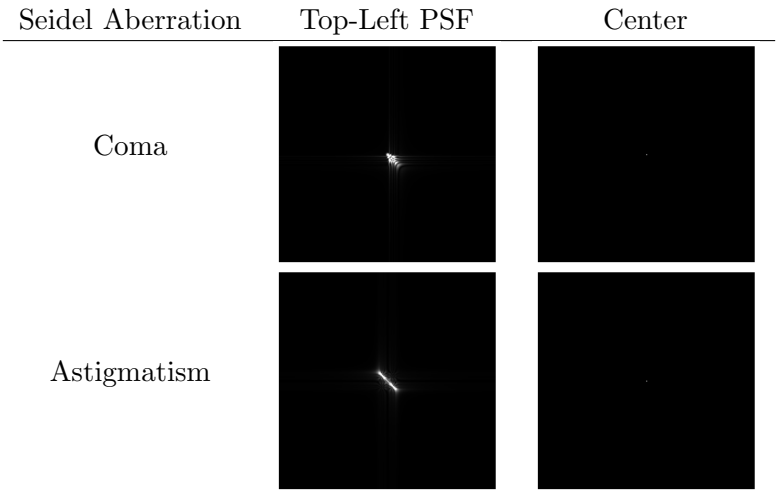

Figure A.1: Coma and Astigmatism Point Spread Functions at different regions of the image

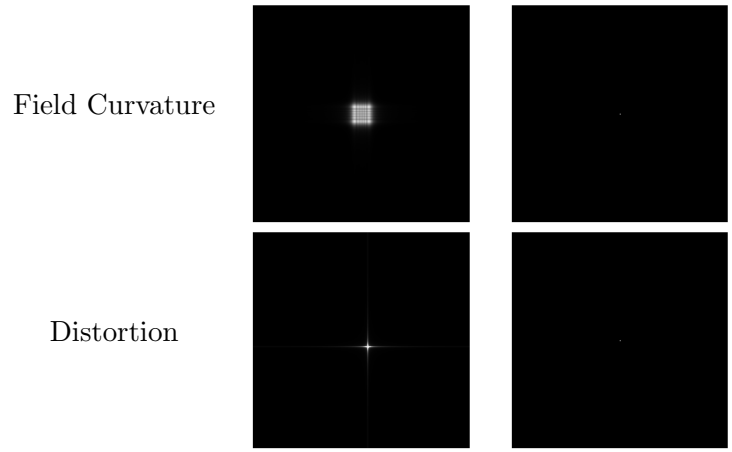

Figure A.2: Field and Distortion Point Spread Functions at different regions of the image

Bibliography

Beck, A. and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *2*(1), 183–202.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.

Chen, L., X. Chu, X. Zhang, and J. Sun. Simple baselines for image restoration.

Dong, J., S. Roth, and B. Schiele (2021). Deep wiener deconvolution: Wiener meets deep learning for image deblurring. *CoRR abs/2103.09962*.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Estatico, C. and F. Di Benedetto. Shift-invariant approximations of structured shift-variant blurring matrices. *62*(4), 615–635.

Fish, D. A., A. M. Brinicombe, E. R. Pike, and J. G. Walker (1995, Jan). Blind deconvolution by means of the richardson–lucy algorithm. *J. Opt. Soc. Am. A 12*(1), 58–65.

Hirsch, M., S. Sra, B. Schölkopf, and S. Harmeling (2010, June). Efficient filter flow for space-variant multiframe blind deconvolution. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 607–614.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks.

Nagy, J. G. and D. P. O'Leary (1998). Restoring images degraded by spatially variant blur. *SIAM Journal on Scientific Computing 19*(4), 1063–1082.

Ng, M. K., J. Koo, and N. K. Bose (2002). Constrained total least-squares computations for high-resolution image reconstruction with multisensors. *International Journal of Imaging Systems and Technology 12*(1), 35–42.

Robbins, G. and T. Huang (1972). Inverse filtering for linear shift-variant imaging systems. *Proceedings of the IEEE 60*(7), 862–872.

Ronneberger, O., P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Voelz, D. G. (2011). Computational fourier optics: a matlab tutorial. *(No Title)*, 51.

Woo, S., S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders.

Yanny, K., K. Monakhova, R. W. Shuai, and L. Waller. Deep learning for fast spatially varying deconvolution. *9*(1), 96.

Zamir, S. W., A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration.

Zhang, A., Z. C. Lipton, M. Li, and A. J. Smola (2023). *Dive into Deep Learning*. Cambridge University Press. https://D2L.ai.

Zhang, K., W. Ren, W. Luo, W. Lai, B. Stenger, M. Yang, and H. Li (2022). Deep image deblurring: A survey. *CoRR abs/2201.10700*.

Zhang, Y., D. Li, X. Shi, D. He, K. Song, X. Wang, H. Qin, and H. Li. KBNet: Kernel basis network for image restoration. version: 1.

Zhang, Y., D. Li, X. Shi, D. He, K. Song, X. Wang, H. Qin, and H. Li (2023). Kbnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881*.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks.