Hierarchically Normalized Models of Visual Distortion Sensitivity

Physiology, Perception, and Application

by

Alexander Berardino

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

May 2018

Eero P. Simoncelli

"I want to stay as close to the edge as I can without going over.

Out on the edge, you see all kinds of things you can't see from the center. ... Big, undreamed-of things – the people on the edge see them first"

– Kurt Vonnegut, Player Piano

Dedication

This thesis is dedicated to the memory of Jonathan Houy and Bobie Bao, whose tireless will, tenacious spirt, and unflappable good humor in the face of extreme adversity left an indelible mark on my soul.

Acknowledgements

I want to begin by thanking my advisor Eero. Eero took a risk on me when I decided to switch from experimental neuroscience to theory after I had already completed a year of my Ph.D., and I hope that risk played out as well for him as it has for me. I have grown so much in these last five and a half years, and his mentorship and support have pushed me to places I never expected. Eero's attention to minute detail, his drive to continually perfect his work, his intellectual rigor, and his willingness to let his mentees spend time exploring ideas that might not pay off, have been an enourmous inspiration to me, and lessons I will take with me throughout my career. I aspire to be half the mentor he is.

I would also like to thank my committee chair, Roozbeh Kiani, for the focus and discipline he brought to my committee, and to the arc of my research. Roozbeh's ability to cut through the fog and drill down to the central question has been invaluable to me during this process. I'd also like to thank my other committe members, Jonathan Winawer, Liam Paninski, and Tony Movshon for their insightful, and often challenging, feedback, inspiring work, and mentorship over the course of the last 5 and a half years. The work presented here would not be half of what it is without their input.

I would like to thank the four people at NYU who most contributed to any success I have had. First, I would like to thank Maureen Hagan, who acted as my first informal mentor throughout the first year of my Ph.D. Maureen was instrumental in helping me to manage the transition into graduate school, and I would not have made it the rest of the way if not for her. Second, I would like to thank Neil Rabinowitz for playing the same role as I transitioned into Eero's lab. Neil is a force of nature who always had time to work through a problem at the blackboard, or to play out a thought experiment with me.

His infectious enthusiasm for the work in the lab convinced me that I had made the right decision, even when I wasn't entirely convinced I could do the work.

Third and fourth, I'd like to thank the two people without whom none of the work presented here would be possible, my collaborators Johannes Ballé and Valero Laparra. Johannes and Valero, I have learned so much from you, from the way you approach a problem, from your work ethic, from your creativity. I am immensely lucky that my time in Eero's lab overlapped with yours. I will take your influence with me for the rest of my career.

Eero's lab is a tremendously collaborative place full of incredibly smart and creative people, and I would be remiss not to thank those who I had the privelage of calling labmates: Corey Ziemba, Andrew Zaharia, Robbe Gorris, Paul Levy, Catherine Olsson, Billy Broderick, Jimmy Wang, Manu Rhagavan, Tim Oleskiw, Laura Noren, Pierre-Etienne Fiquet and Carloline Haimerl. I'd like to also thank the many students who rotated through the lab throughout my five and half years: Long Sha, Reuben Feinman, Nikhil Parthasarathy, Bas Van Opheusden, and Colin Bredenberg. I'd like to especially thank Olivier Henáff, my office mate and frequent sounding board, for being a continuous source of inspirtation and good humor.

I would also like to thank my collabotor David Brainard, from the University of Pennsyvlania, who is always at the ready to answer my unending string of questions and confusions, as well as Jonathan Shlens, from Google Brain, for his helpful and insightful feedback on my work, and for his personal encouragment and advice.

Several exceptional NYU undergraduates collected the majority of the data presented in this thesis: Natalie Pawlak, Rebecca Walton and Lydia Cassard. I know that you often spent your weekends collecting data for my projects and I am extremely thankful for your

hardwork and dedication.

There are, unfortunately, too many people at CNS to thank each individually. But I would like to thank Luke Hallum and Najib Majaj for their helpful feedback, guidance and mentorship. Thank you Lynne Kiorpes and Paul Glimcher for your invaluable wisdom and the lessons I learned from you about teaching. Thank you Weiji Ma for your indomitable public spiritedness. Thank you Heather, Gary, Weiji, Steve, Catherine, Huayi and Li, for helping me found, manage, and carry on NeuWrite Downtown. Thank you Gerick, for making me laugh. Thank you Aaron, for watching improvised song-writing shows in Washington Square with me. And finally, thank you Amala, Jess, Keith, Erik, Jenna and the entire administrative staff for making my time here so seamless and enjoyable.

I have made many friends in my time at CNS, but none as close as my classmates. Thank you, Emily, Kristina, Brittany, Oliver, Christina, Robert, Aina, Evelyn, Evan, Andra and Carolina. I will miss our friendsgivings tremendously.

Finally, I would like to thank my family. To my parents, thank you for always supporting my passions, even when they didn't make sense to you, for letting me go so far away from you in pursuit of my dreams, and for supporting me as I am in every way. Thank you also for buying me new coats when I come home from New York in tatters. To my brother Stephen and sister-in-law Jordyn, thank you for bringing me two incredible nephews, the joy they bring me has gotten me through some of the tougher times in my Ph.D. To my aunts, uncles and cousins, thank you so much for your continuous love and support, and for being entertained by my stories about monkeys. And to my partner, Jason, I don't know where I would be without you. I am so lucky to have found a person who enthusiastically supports my passions and my pursuit of my goals. You are my source of strength and joy and I am immensely grateful for you.

Preface

Chapters 1,2, and 3 were the result of a close collaboration with 2 postdoctoral research fellows, Johannes Ballé and Valero Laparra, in the lab of Eero Simoncelli. Most of this work has been published. Chapter 4 represents an ongoing collaboration with David Brainard at the University of Pennsylvania and is unpublished. Ponomarenko et al., (2009)

Abstract

How does the visual system determine when changes to an image are unnatural (image distortions), how does it weight different types of distortions, and where are these computations carried out in the brain? These questions have plagued neuroscientists, psychologists, and engineers alike for several decades. Different academic communities have approached the problem from different directions, with varying degrees of success. The one thing that all groups agree on is that there is value in knowing the answer to the question. Models that appropriately capture human sensitivity to image distortions can be used as a stand in for human observers in order to optimize any algorithm in which fidelity to human perception is necessary (i.e. image and video compression).

In this thesis, we approach the problem by building models informed and constrained by both visual physiology, and the statistics of natural images, and train them to match human psychophysical judgments about image distortions. We then develop a novel synthesis method that forces the models to make testable predictions, and quantify the quality of those predictions with human psychophysics. Because our approach links physiology and perception, it allows us to pinpoint what elements of physiology are necessary to capture human sensitivity to image distortions. We consider several different models of the visual system, some developed from known neural physiology, and some inspired by recent breakthroughs in artificial intelligence (deep neural networks trained to recognize objects within images at human performance levels). We show that models inspired by early brain areas (retina and LGN) consistently capture human sensitivity to image distortions better than both the state of the art, and better than competing models of the visual system. We argue that divisive normalization, a ubiquitous computation in the visual system, is

integral to correctly capturing human sensitivity.

After establishing that our models of the retina and the LGN outperform all other tested models, we develop a novel framework for optimally rendering images on any display for human observers. We show that a model of this kind can be used as a stand in for human observers within this optimization framework, and produces images that are better than other state of the art algorithms. We also show that other tested models fail as a stand in for human observers within this framework.

Finally, we propose and test a normative framework for thinking about human sensitivity to image distortions. In this framework, we hypothesize that the human visual system decomposes images into structural changes (those that change the identity of objects and scenes), and non-structural changes (those that preserve object and scene identity), and weights these changes differently. We test human sensitivity to distortions that fall into each of these categories, and use this data to identify potential weaknesses of our model that can be improved in further work.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Quantifying the Visibility of Image Distortions

Digital images are subject to many potentially corrosive processes that introduce artifacts, deviations, and distortions in the course of image capture, compression, transmission and reproduction. Human observers can easily identify distortions, and can classify deviations that lead to images that appear unnatural (and thus degrade the quality of the image) separately from those that are natural (and do not degrade the quality of the image, or even improve it). For many years, the field of image quality assessment (IQA) has been structured around studying the set of distortions that are encountered in the processing of digital images, and human sensitivity to them. The goal of this field has been to build a metric, or model, that quantifies the visibility of different types of image distortions, which can be utilized to benchmark algorithms for all of the above processes in the image capture and display pipeline.

There are three main approaches to this problem. The first, known as *full-reference*, assumes a ground truth image is available for comparison to the distorted image. The second, *no-reference*, works with only the distorted image, and compares the image to a

model of natural images (see Ma. et al., (Aug. 2017) and Ma et al., (2018) for examples of modern no-reference IQA algorithms). The third, *reduced-reference*, assumes that a reference image is only partially available, through extracted features or statistics. While the latter two approaches have appealing applications both within engineering and for understanding neural systems, the work presented in this thesis will focus entirely on the first approach, full-reference IQA. This form of the problem aligns most closely with the types of experiments we will carry out in the lab, as well as with the types of applications we will pursue.

In the full-reference case, we can think of distorted images as an original image, X, plus some corrupting unit vector, $\vec{u}$ multiplied by a scalar amplitude, α. The simplest, and for a long time most-common, choice to attempt to quantify perceptual sensitivity to distortions was to use the mean squared error (MSE) between the pixel values of the original ground truth image, X, and the distorted version of that image, $\hat{X}$.

$$D(X, \hat{X}) = ||X - \hat{X}||_2 = ||(\alpha\vec{u})^2|| = \alpha$$

In this framework, larger MSE between two images indicates more visible distortion. Under the assumptions above, MSE recovers the unit distortion vector, $\vec{u}$, as well as the amplitude value, α, but only retains the amplitude as a measure of distortion. This model implicitly assumes that the only attribute of a distortion that matters for assessing its visibility is the amplitude, α, and that the type of distortion, or the direction that the vector $\vec{u}$ points in the space of possible distortions, is not relevant for assessing visibility. While the amplitude of the distortion is certainly important, a simple demonstration reveals why reliance on amplitude alone is not sufficient to capture perceptual sensitivity to image distortions. Figure 1.1 shows 8 images which have equivalent MSE when compared to the

Figure 1.1: **Failures of MSE as a Perceptual Metric**: Each of the 8 images surrounding the center image has the same MSE when compared to the center image. Despite this, the images are of very different levels of degradation, showing that MSE does not capture perceptual sensitivity to image distortions well. (Adapted from (Ponomarenko et al., 2009))

image in the center, or put differently, the same amplitude of distortion. It is immediately apparent that all 8 images are not of equivalent perceptual quality. In fact some of the images are quite degraded, while others appear nearly identical to the original. This result has been confirmed by several careful studies of visual quality assessment (Eckert & Bradley, (Nov. 1998), Eskicioglu & Fisher, (Dec. 1995), Girod, (1993), Teo & Heeger, (1994b), Wang & Bovik, (Mar. 2002), Wang, (Dec. 2001), Winkler, (1999), and Z. Wang & Lu, (May 2002)).

MSE fails because it implicitly assumes that humans make judgments about image distortion by comparing the pixel values as rendered on the screen. In reality, however, the brain doesn't have access to the raw pixel values. Instead, to make this judgment, human subjects are comparing representations of those pixel values within neural populations within their visual system. Distortions that are of equivalent amplitude in raw pixel differences may not be of equivalent amplitude, or discriminability, in their neural representation. The failure of MSE as a metric has led many to realize that an appropriate IQA metric has to be sensitive to both the amplitude, α, and the direction, $\vec{u}$, of the distortion, and has to re-weight its sensitivity along different dimensions to accommodate the fact that humans are differently sensitive to distortions of equal amplitude that lie in different directions (see figure 1.2).

There are many approaches to solve this problem, the simplest and most appealing of which is to build a function, f, that approximates the transformations of the human visual system, and to take the difference between the representations of $\vec{x}$ and $\hat{x}$ within the representation space of f.

$$D(X, \hat{X}) = ||f(X) - f(\hat{X})||_2$$

This approach was first pursued by Mannos and Sakrison in 1974 in their pioneering study "The Effects of a Visual Fidelity Criterion on the Encoding of Images" (Mannos & Sakrison, 1974). This approach was subsequently followed up on by many researchers, and the functions they created to approximate the human visual system quickly became more detailed, informed by troves of psychophysical and physiological data describing the human visual system that was being concurrently amassed(see Teo & Heeger, (1994b) and Watson, (Jan 2000) and Eckert & Bradley, (Nov. 1998) and Pappas & Safranek, (2000) for reviews.

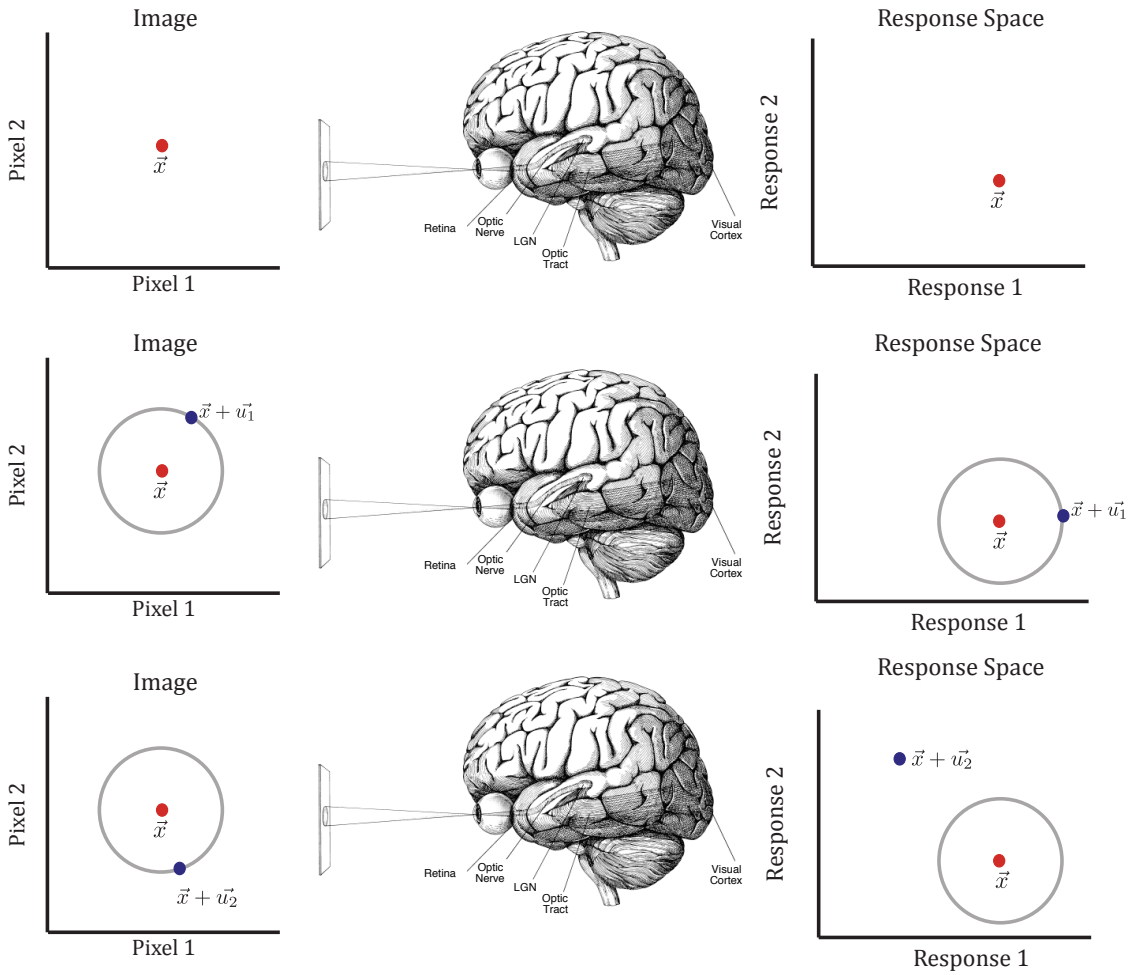

Figure 1.2: **Humans are differently sensitive to distortions along different dimensions.** We can plot a hypothetical 2 pixel image, $\vec{x}$, as a point in a 2 dimensional space. Around this point, there is a circle of distorted images, $\hat{x} = \vec{x} + \vec{u}$, such that all $\hat{x}$ have the same amplitude, α and MSE. Human judgments about the similarity of $\vec{x}$ and $\hat{x}$ are computed based on the similarity of the representations of $\vec{x}$ and $\hat{x}$ within noisy neural populations within their visual system (here plotted as noise clouds representing a set of neural responses upon repeated viewing). Human sensitivity to a particular distortion is limited by the amount of overlap between the noisy response to the original image, and the noisy response to the distorted image. Despite the fact each of the distortions, $\vec{u}_1$ and $\vec{u}_2$, had the same amplitude in the pixel space, they may have very different amounts of overlap in this neural space after having been transformed by they human visual system, and thus the human viewer will be differently sensitive to each of them. In the example pictured above, the observer is more sensitive to changes in the direction of $\vec{u}_2$ than $\vec{u}_1$.

Despite this complexity, a study by the Visual Quality Experts Group released in March 2000 showed that the performance of most models were statistically indistinguishable from PSNR (Peak Signal-to-Noise Ratio), a direction agnostic quality metric computed from the ratio of the maximum displayable luminance to the MSE (VQEG, (Mar. 2000)).

$$PSNR(\vec{x}, \hat{x}) = 10 * log_{10}\Big(\frac{I_{max}^2}{MSE(\vec{x}, \hat{x})}\Big)$$

Where I_{max} is the maximum displayable luminance (or pixel value).

This study called into question the entire approach to building a cohesive model of the action of the human visual system on natural images from the bottom up , and cleared the field for a different approach, the top-down approach, to step in and become the defacto standard. In 2004, Wang et al. introduced the Structural Similarity Metric, based on the hypothesis that the human visual system is highly adapted to extract structural information from the world (Wang et al., (2004)). Additionally, they proposed that instead of trying to model the action of the visual system by appending the results from many psychophysical studies, you could instead use this intuition to design a metric that separates structural and non-structural elements of images, and weighs them differently (See Figure 1.3). SSIM worked significantly better than PSNR, and all of its contemporary competitors, and quickly became the defacto standard visual quality metric, even winning a Engineering Emmy for contributions that significantly improve existing methods, or innovations that materially affect the transmission, recording or reception of television in 2015.

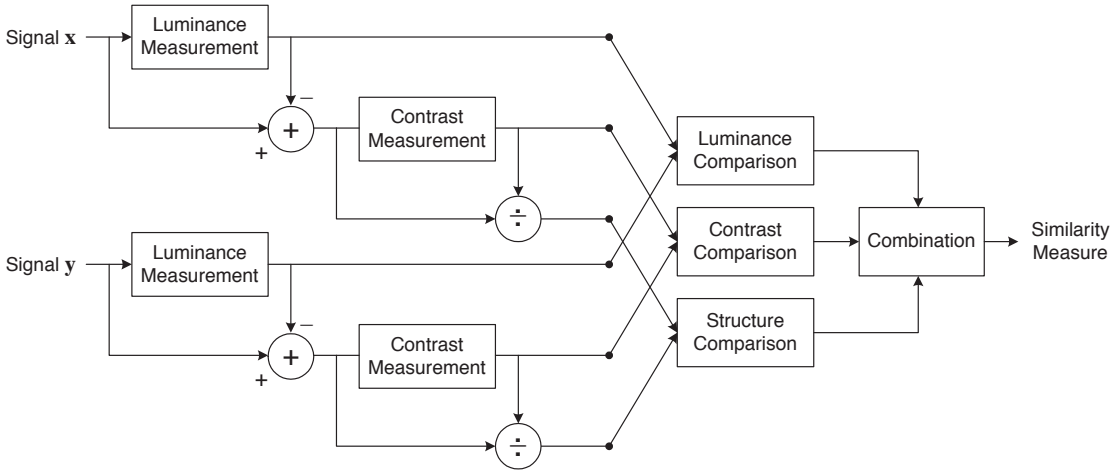

Figure 1.3: SSIM Signal Processing Diagram: SSIM decomposes both images into a measure of local luminance, a measure of local contrast, and a measure of local structure. In order to compare the two images, SSIM finds the correlation between images for each of these measures. Finally, these three measures are exponentiated and combined by taking the product between them. Adapted from Wang et al., (2004)

1.2 Returning to the Bottom-Up Approach with Better Neural Models

Despite its success, it is difficult to relate SSIM directly to functions carried out in actual neural networks, leaving open questions about how to relate it to neuroscience as well as how best to extend its function. In addition, recent advances in modeling neural responses have called into question the fundamental assumptions that led to the abandonment of the bottom-up approach in 2003 that paved the way for SSIM's top-down approach.

In their paper, Wang et. al laid out what they saw as the four main issues with the traditional bottom up approach that can be reduced down to two fundamental points Wang et al., (2004). The first problem arises from the fact that the majority of the models were constructed out of linear, or quasi-linear, operators created to capture psychophysical phe-

nomena that were characterized on simple unnatural stimuli under restricted experimental paradigms. Despite reproducing the observed psychophysical results, it was not clear that these models generalized from simple stimuli to natural images, or from restricted experimental settings to the much larger space of real-world conditions. The second problem arises from the fact that the majority of these models utilized a distance metric, like MSE in their response space, which implicitly assumes that errors in different spatial locations are independent. This assumption stands in contradiction to the inherent correlation of neighboring pixels within images, and most of the linear and quasi-linear decompositions employed in these models did not effectively reduce the correlation between channels, and thus were combining redundant error signals.

Since this time, there have been significant advances in modeling neural responses to natural images, many of which answer directly to the above issues. In this thesis, we aim to utilize these advances to return to the bottom-up approach, and attempt to relate the structure of visual physiology to the perceptual problem of image quality assessment. We will examine several different models of neural physiology, each of which is inspired by different approaches to modeling neural populations.

1.3 LN Models and Deep Neural Networks

Neurons have long been modeled as a combination of linear filters or receptive fields (approximating dendritic summation) and pointwise nonlinearities (approximating spiking nonlinearities) (see Chichilnisky, (2001), Enroth-Cugell & Pinto, (1970), Marmarelis & Naka, (1972), Marmarelis, (1978), Movshon et al., (1978), and Simoncelli et al., (2004) for early examples, reviews, and methods for estimating model parameters from neural data). These models, known as LN models, provide a simple, mathematically tractable approxi-

mation to more biologically detailed models of single neurons, such as the Hodgkin-Huxley model, that capture the functional behavior of neurons (Hodgkin & Huxley, 1952). Simple LN models, as well as short cascades of LN operations, have been shown to explain responses of neurons in many sensory areas.

In 1980, Kunihiko Fukushima introduced the Neocognitron, a stacked LN model, or a multilayered artificial neural network in the terminology of artificial intelligence, based on the work of Hubel and Wiesel explaining their observations of cell responses in the first cortical visual area, V1 (Fukushima, (1980) and Hubel & Wiesel, (1959)). Neocognitron was capable of very simple handwritten character recognition and served as a proof of concept that stacked LN models could perform complicated perceptual tasks (Fukushima, 1980). In 1991 Kurt Hornik published a proof that neural networks of this form (Mulitlayer feedforward networks) could universally approximate, up to a given level of error, any continuous function (Hornik, (1991)). This led to great excitement about the potential of neural networks as a tool for artificial intelligence, but a lack of computational power, data, and difficulty learning appropriate parameters led early neural networks to fade into obscurity (LeCun et al., 2015). The combination of several breakthroughs have resuscitated neural networks in recent years: massive increases in compute power, the emergence of big data, the invention of convolutional neural networks (neural networks that share the same weights at every spatial location), rediscovery of the power of old training methods (such as stochastic gradient descent and backpropogation), and the recognition that deeper networks, or networks with more layers of LN operations, could learn to solve more complicated and abstract tasks than shallow networks (LeCun, 1985; LeCun et al., 1989; Rumelhart et al., 1986 and see LeCun et al., 2015 for a review of the reemergence of deeper neural network methods). This new version of artificial neural networks, known as deep

neural networks, burst onto the scene in 2012 with the introduction of AlexNet, a deep neural network that competed in the annual ImageNet Large Scale Visual Recognition Challenge, a 1000-way object recognition challenge (Krizhevsky et al., 2012). The network not only won the competition and set new performance records, but effectively redefined what was achievable, with a top-5 error more than 10 percentage points ahead of the next best algorithm (Krizhevsky et al., 2012). In the years hence, deep neural networks have been utilized to solve many complicated, and previously unsolved, sensory and cognitive tasks, and are now the defacto solution for many problems in artificial intelligence (LeCun et al., 2015).

Since 2013, several groups have shown a striking similarity between responses in deep layers within performance optimized deep neural networks and the responses of neurons in deep areas of the ventral visual stream, such as V4 and IT, that had previously been difficult to model (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Yamins et al., 2013).This approach to modeling neural circuits not by fitting to neural data, but by forcing neural circuits to perform a complicated objective at human levels by was introduced by Yamins and DiCarlo in 2014 and expanded on in their 2016 review paper, "Using goal-driven deep learning models to understand sensory cortex" (Yamins & DiCarlo, (2016)). This strategy is based in the principle that a neural network will have to be effective at solving the behavioral tasks the sensory system supports to be a correct model of a given sensory system (Yamins & DiCarlo, (2016)). In their work, they first train the parameters of a neural network model, constrained to loosely reflect the architecture of the ventral visual stream, to perform an ethologically relevant task, in their case a complicated form of object recognition (See Figure 1.4). They subsequently compare the optimized network to neural responses and find that they are able to predict neural responses in macaque

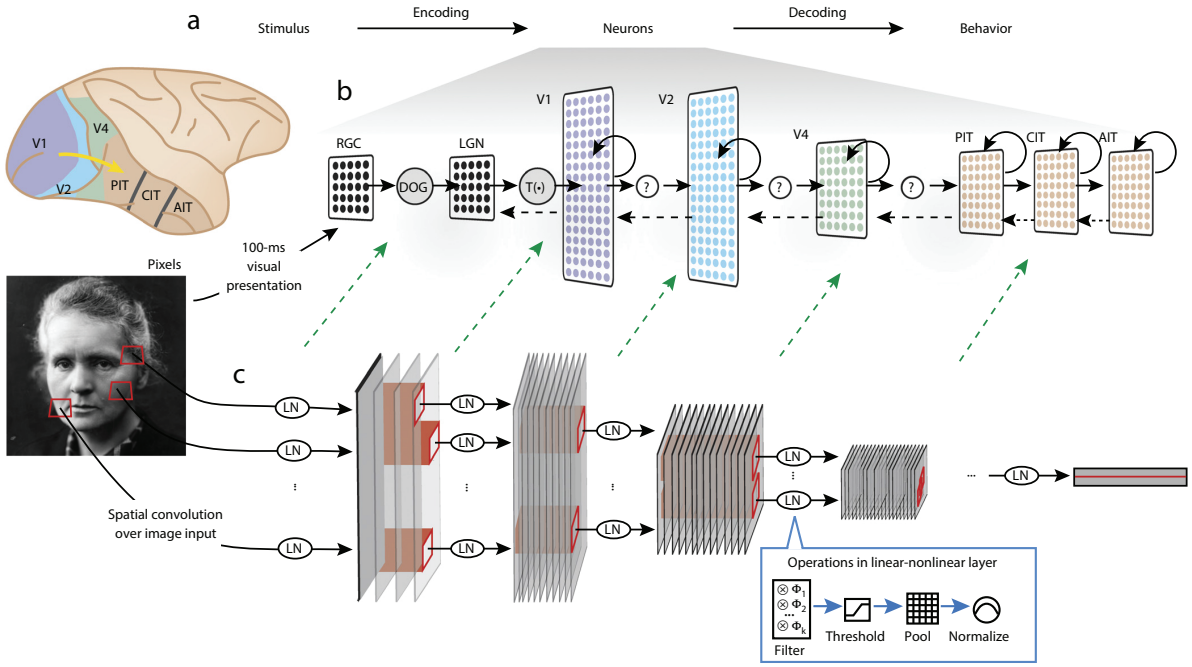

Figure 1.4: **Similarity of Deep Neural Network Architecture and the Ventral Visual Stream**: The architecture of deep neural networks trained to perform complicated visual tasks at or above human performance levels superficially resemble the architecture of the primate ventral visual stream. (Adapted from Yamins & DiCarlo, (2016))

IT and V4 better than any previous image computable models (Yamins et al., 2014) (See Figure 1.5). Several other groups have subsequently shown that these neural networks also predict fMRI responses along the ventral stream in human subjects in response to natural images(Cichy et al., (2016) and Khaligh-Razavi & Kriegeskorte, (2014)). Additionally, it has been shown that despite the fact that these networks are trained only on a global objective computed from the output of their latest layers, the filters in early layers of these networks resemble localized oriented band pass filters, like those found V1, and early layers of these networks predict fMRI responses in early ventral visual stream areas in humans (Khaligh-Razavi & Kriegeskorte, (2014), See Figures 1.5 and 1.6).

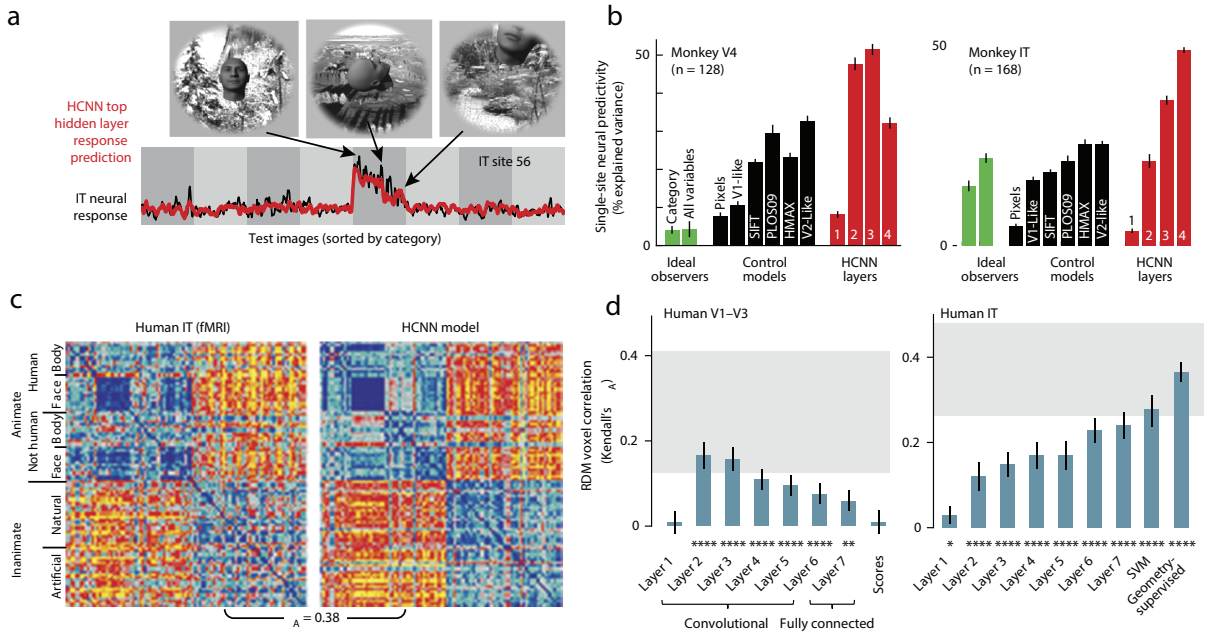

Figure 1.5: **Deep Neural Networks Trained On Complex Visual Tasks Predict Responses Along the Ventral Visual Stream.** Responses in late layers of deep neural networks are predictive of responses of V4 and IT neurons to natural images (A and B). In addition, the metric across object classes in human IT and the response space of neural networks show a large degree of correlation (C). Finally, early layers of performance-optimized deep networks predict fMRI activity in early layers of the ventral visual stream (D). (Adapted from Yamins & DiCarlo, (2016))

Figure 1.6: Visualizing Filters from the first layer of AlexNet, a deep neural network trained on object recognition. Learned filters from the first layer are localized, oriented band pass filters resembling response properties of neurons found in the first cortical visual area, V1. (Adapted from Krizhevsky et al., 2012)

The success of these networks at capturing complicated neural and behavioral responses to natural images suggests that they may be sufficiently nonlinear, and sufficiently good models of visual processing in realistic situations, to overcome the limitations of earlier linear and quasi-linear models used for image quality assessment. In fact, several authors have already begun exploring this space, constructing distortion metrics from the responses of intermediate and deep layers of several different deep neural networks (Dosovitskiy & Brox, (2016), Hénaff & Simoncelli, (2016), Johnson et al., (2016), and Parthasarathy et al., (2017)). Many of these authors have employed these metrics as perceptual loss functions in order to optimize a neural network to generate images under different constraints, and have shown that networks optimized with these perceptual loss functions produce better results than networks optimized with MSE (See Figure 1.7 for a result from Johnson et al., (2016)). Though these results are potentially exciting, no one has rigorously explored the capability of deep neural networks to capture human perceptual sensitivity, nor compared their performance to other models that outperform MSE. In this thesis, we will take up this objective.

1.4 Moving Beyond LN Models

1.4.1 Coding Efficiency

While deep neural networks may address some of the concerns Wang et al. expressed about the bottom up approach, it is not clear that a cascade of LN operations addresses their redundancy concerns (Wang et al., (2004)). As such, we would like to compare the performance of stacked LN operators to a set of models designed to simultaneously match neural responses to natural stimuli, and reduce redundancy between output coefficients.

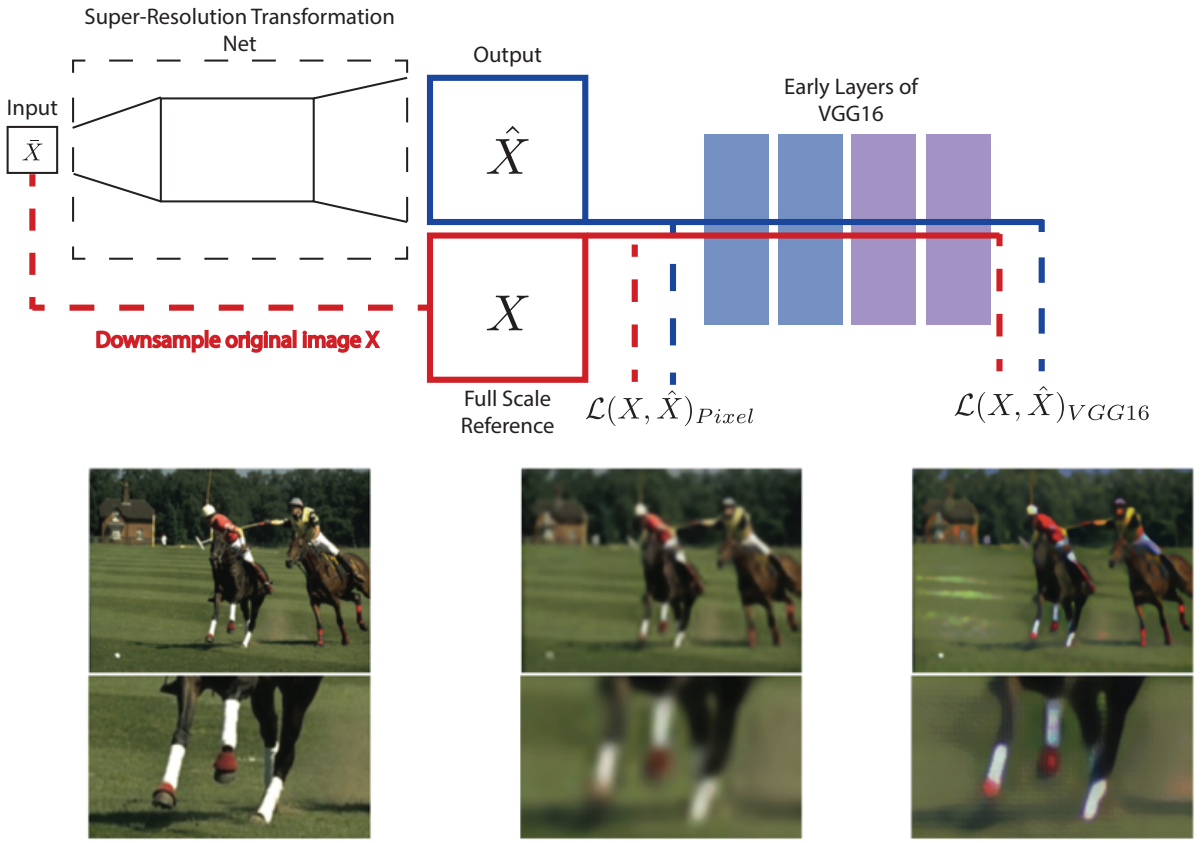

Figure 1.7: **Deep Neural Network based loss functions for Super-Resolution**: Johnson et. al. trained two neural networks to produce $x4$ and $x8$ super-resolution images from down-sampled inputs. In order to train the weights of the networks, the authors computed the perceptual distortion distance between the original, full-size images, and the super-resolution images that the network produced, and backpropogate the errors to the weights of the network. The first network was trained using Pixel MSE as a perceptual metric. The second network was trained using MSE computed within the response space of intermediate layers of VGG16, a deep neural network trained on object recognition (Simonyan & Zisserman, (2015)). The authors show that across many images, the network trained to minimize the VGG16-based metric produces significantly better images than the network trained with Pixel MSE. (Adapted from Johnson et al., (2016))

In 1961, Horace Barlow hypothesized that the goal of neural systems (specifically early vision) was to remove statistical redundancy in natural signals it encountered (Barlow, 1961; Simoncelli & Olshausen, 2001). This is known as the efficient coding hypothesis. In 1996, Olshausen and Field showed that optimizing a linear filter bank from natural images for a sparsity objective (a variant of the efficient coding hypothesis) produced filters that resemble V1 neurons (much like the first stage of deep neural networks)(Olshausen & Field, (1996)). Following this finding, a class of methods oriented around searching for a linear representation of images (or data more generally) in which the output coefficients are as independent as possible, known collectively as Independent Components Analysis (ICA), arose (Hyvärinen & Oja, 2000). It was similarly shown that this objective led to basis functions that resemble V1 simple-cells (Bell & Sejnowski, 1997; Hyvärinen & Oja, 2000). These results suggest that models of V1 simple-cells would be a good place to start if one wanted to simultaneously match neural responses to natural stimuli, and reduce redundancy between output coefficients.

In the early 2000's, however; Schwartz and Simoncelli showed that even though ICA filters were optimized to make signals as independent as possible, there were still significant higher-order correlations between filter responses to natural images (See Figure 1.8) (Schwartz & Simoncelli, (2001)). Schwartz and Simoncelli also showed that a simple nonlinear operation, divisive normalization, could easily remove these remaining redundancies, producing signals that were significantly more independent (See Figure 1.8) (Schwartz & Simoncelli, (2001)). Divisive normalization, in which the response of a neuron is divided by a weighted sum of the responses of its neighbors, has been found throughout the brain and has been hypothesized to be a canonical neural computation (Carandini & Heeger, (2012)).

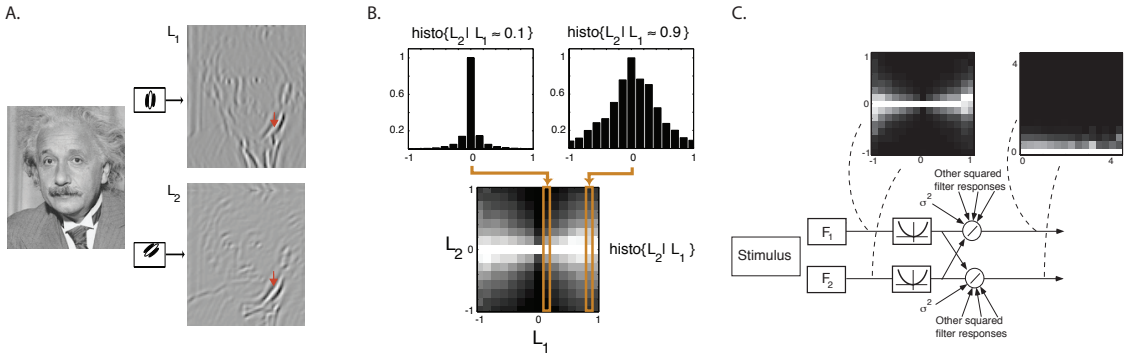

Figure 1.8: **Normalization Reduces Correlated Variability of Linear Filter Outputs**: Schwartz and Simoncelli examined the higher order correlations between the outputs of 2 linear bandpass filters designed to make image signals independent (A). By examining the joint histograms of 2 filters at different spatial locations within an image, they found that the filters displayed higher order correlations (specifically, while the responses of L1 and L2 are decorrelated, the variance of the distribution of L2 increases with increasing value of L1) (B). Finally, they showed that by dividing the linear filter responses by a weighted sum of the squared responses of neighboring filters produced responses that were truly independent (C). (Adapted from Schwartz & Simoncelli, (2001))

These results together suggest that the addition of a simple divisive nonlinearity, changing the LN model into an LG model, may provide a powerful toolset for simultaneously reducing redundancy in model representations and better capturing responses of real neural circuits with fewer layers. This finding is no doubt partially responsible for the success of SSIM, which implements a form of divisive normalization, as well as the more recent success of models of V1 containing normalization between filters as perceptual metrics (Laparra et al., (2010)).

Models like that of Laparra et al., (2010), as well as traditional solutions to the efficient coding hypothesis such as ICA, curiously bypass early processing stages in the retina and LGN and begin with a bank of linear filters that resemble the function of V1. However, in 2011, Karklin and Simoncelli showed that under realistic assumptions of metabolic con-

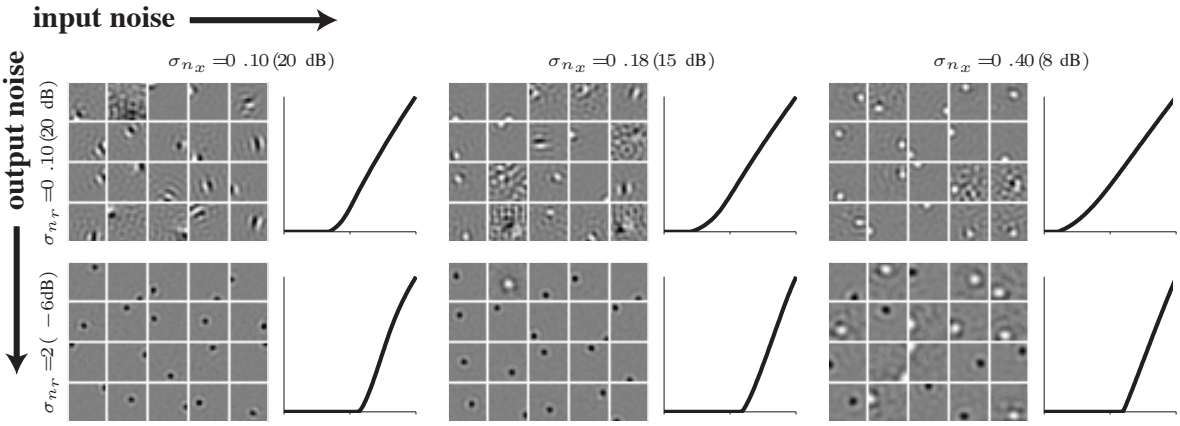

Figure 1.9: **On-Off filters Optimally Reduce Mutual Information in the Presence of Noise**: Karklin and Simoncelli showed in 2011 that filters optimized for efficient coding under the presence of significant input and output noise (matching real world conditions) did not produce V1-like oriented filters (as in the case with no noise), but instead produced populations of On and Off center-surround filters like those found in the earliest visual processing stages, the retina and LGN. (Adapted from Karklin & Simoncelli, (2011))

straints, as well as the presence of input (cone) noise and output noise, the efficient coding framework produces a population of On and Off center-surround filters, like those found in the Retina and LGN (See Figure 1.9)(Karklin & Simoncelli, (2011)). Additionally, it has long been known that the responses of cells in the Retina and LGN are highly non-linear, and their function cannot easily be subsumed by a simple linear filter (See Figure 1.10) (Mante et al., (2005), Mante et al., (2008), and Shapley et al., (1972)). For example, in 2008, Mante and Carandini showed that they could capture a large percentage of the response variance of cat LGN neurons in response to drifting gratings and natural images using an LN model with hierarchically stacked stages of gain control following a fixed linear filter (See Figure 1.11)(Mante et al., (2008)). Thus, under realistic physiological noise assumptions, an LG model of On and Off center-surround filters followed by successive stages of divisive normalization (which is a static version of the dynamic gain control from Mante

et al., (2008)) both reduces correlations between output coefficients more than a model of oriented band-pass v1-like filters and captures neural responses of the first stages of visual processing in the primate visual system.

In this thesis, we will build on these observations, and test the ability of a model of the often-overlooked early visual processing (Retina and LGN), with hierarchically cascaded stages of divisive normalization inspired by the model of Mante and Carandini, to capture human perceptual similarity (Mante et al., (2008)). We will compare the performance of this highly nonlinear one stage model to the performance of much deeper neural networks constructed from simple LN operators, as described above. To do so, we will draw inspiration from the goal-directed network modeling of Yamins and DiCarlo, and fit the parameters of each of our physiological models such that they perform a natural image-based psychophysical task at a high level. Details of the models, task, data and optimization are described below.

1.5 Capturing Perceptual Distortion Sensitivity within Neural Networks

In order to capture human perceptual distortion sensitivity within our set of neural network models and to fit the parameters of the models that are unknown, or underconstrained by physiological data, we trained each of the networks to predict human sensitivity to distortions to natural images. We utilized the publicly available TID-2008 database, which contains a large set of distorted image pairs and corresponding human ratings of distortion (Ponomarenko et al., (2009)).

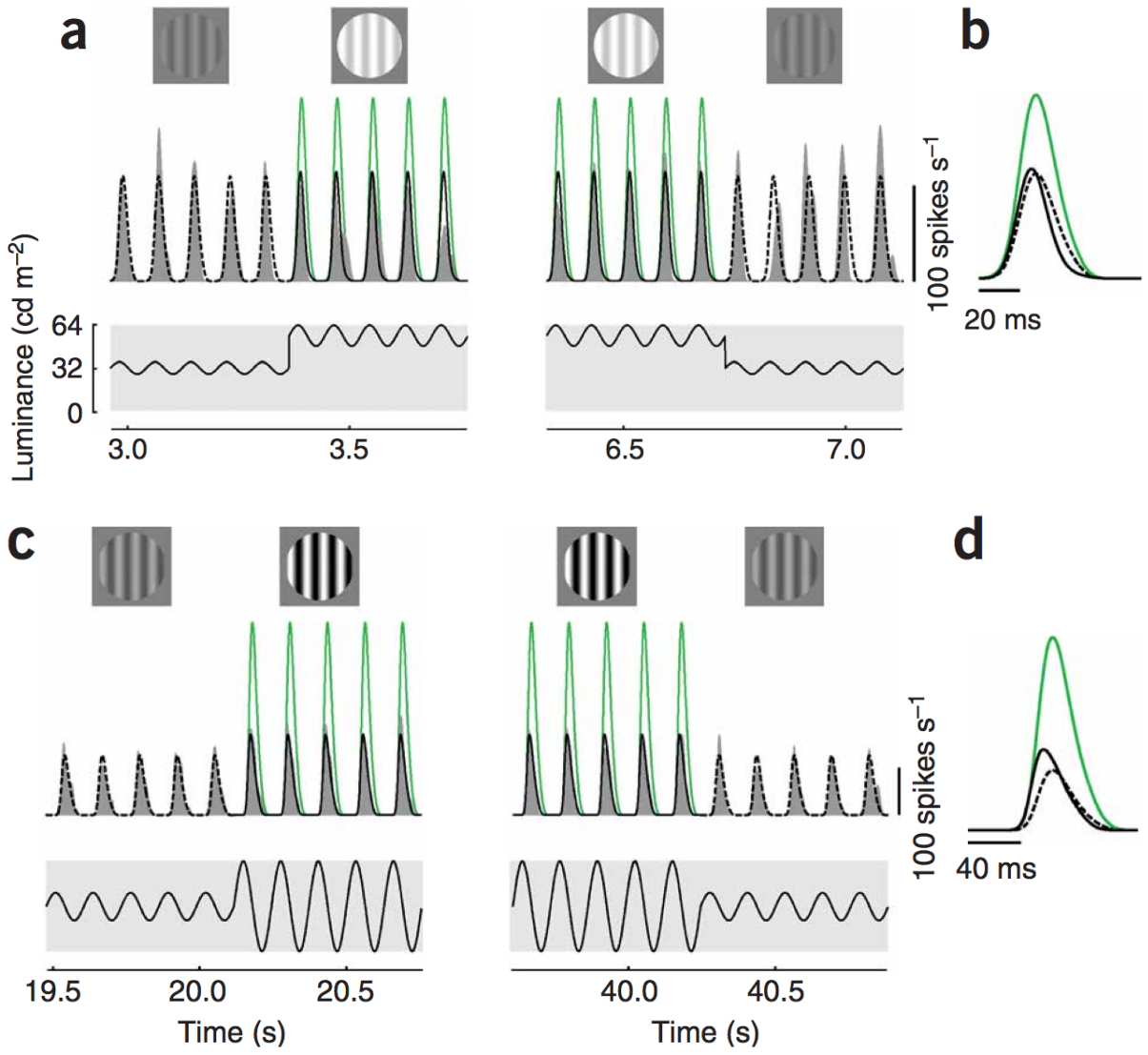

Figure 1.10: **Nonlinear response of Cat LGN to a step in Luminance and Contras**: Predictions of a linear filter (In green) and measured neural responses from cat LGN in response to: (A.) an increment in luminance, (B.) a decrement in luminance, (C.) an increment in contrast, and (D.) a decrement in contrast. In each case, the linear filter poorly predicts the behavior of the neurons. (Adapted from Mante et al., (2005))

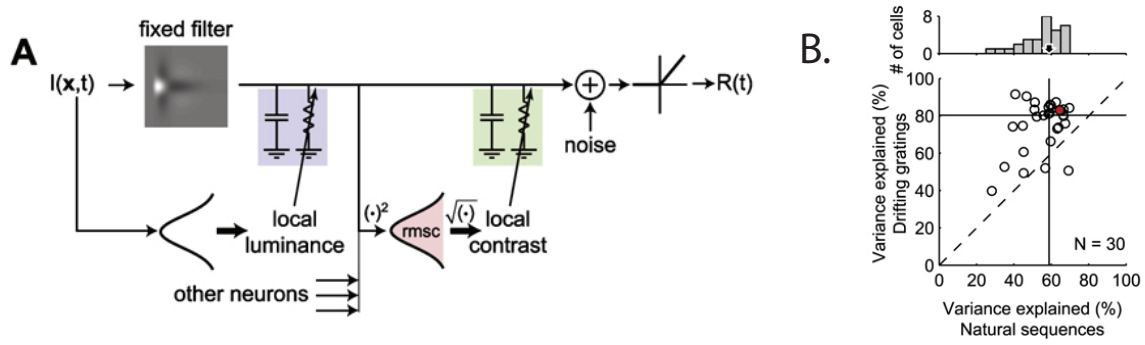

Figure 1.11: A dynamic normalized model of Cat LGN explains a large fraction of response variance to drifting gratings and Natural Images. (Adapted from Mante et al., (2008))

1.5.1 Estimating Model Parameters from Perceptual Data

Perceptual distortion distance for each model was calculated as the Euclidean distance between the model's representations of the original, $\vec{x}$, and distorted images, $\vec{x}\,'$:

$$D_\phi = ||f_\phi(\vec{x}) - f_\phi(\vec{x}\,')||_2$$

For each of our models, we optimized the parameters, ϕ, so as to maximize the correlation between the model-predicted perceptual distance, D_ϕ and the human mean opinion scores (MOS) reported in the TID-2008 database:

$$\phi^* = \arg\max_\phi \Big(\text{corr}(D_\phi, MOS)\Big)$$

Optimization for all models was performed using regularized stochastic gradient ascent with the Adam algorithm and backpropogation, with the exception of our deepest neural network, for which optimization was performed using non-negative least squares (details below).(Kingma & Ba, (2014)).

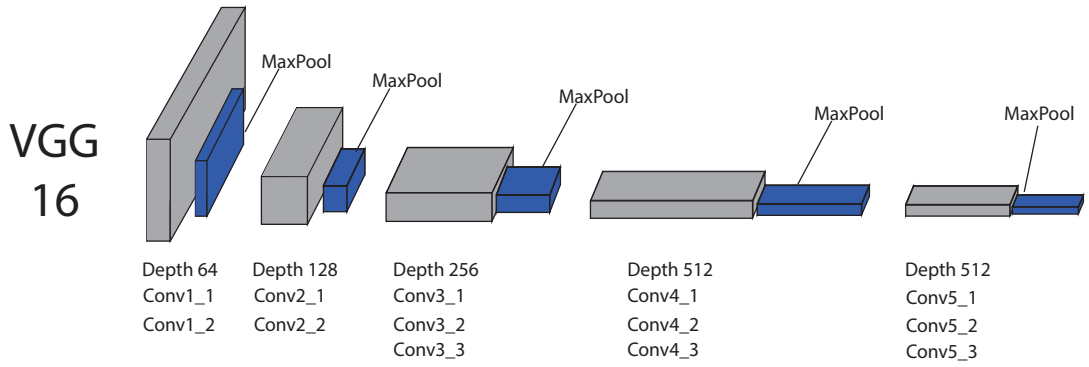

Figure 1.12: **Architecture of VGG16**: VGG16 is constructed from a series of convolutional filtering layers, rectified linear units (ReLu), and max-pooling stages. We only consider the first thirteen layers of the network here (those preceding the fully-connected layers) (Adapted from Simonyan & Zisserman, (2015))

1.5.2 Deep Neural Networks

VGG-IQA

We begin with one of the deep neural networks most commonly utilized as a perceptual metric, VGG16 (See figure 1.12) (Simonyan & Zisserman, (2015)). We call our modified version targeted at image quality assessment (VGG-IQA). VGG16 performs object recognition at human levels, and we want to preserve the information about vision that VGG16 has learned in the process of being optimized for this task. As such, our version leaves the linear filters trained on object recognition as they are, and simply computes a weighted mean squared error over all rectified convolutional layers of the VGG16 network between the image pairs (13 weight parameters, ψ, in total):

$$D(\vec{x}, \vec{x}\,')_{vgg} = \sum_{i}^{13} \psi_i ||layer_i(\vec{x}) - layer_i(\vec{x}\,')||_2$$

with weights trained on the perceptual task above. The output space created by appending each layer of VGG16 is remarkably large, and creates problems for optimizing our parameters using stochastic gradient descent. Instead, we precomputed the layer-wise MSE between each image pair in our database, and utilized non-negative least squares to optimize the 13 weight parameters, ψ, to best match the reported human distortion ratings.

4-stage Cascaded LN Network

VGG16 is a powerful neural network capable of solving complicated visual tasks. However, in utilizing a network pre-trained on a separate task, we are not taking advantage of the full power of neural networks to learn from data. To rectify this, we constructed a generic 4-layer convolutional neural network (CNN, 436908 parameters - Fig. 1.13) which we will train from scratch on the perceptual task described above.

Within this network, each layer applies a bank of 5×5 convolution filters to the outputs of the previous layer (or, for the first layer, the input image). The convolution responses are subsampled by a factor of 2 along each spatial dimension (the number of filters at each layer is increased by the same factor to maintain a complete representation at each stage). Following each convolution, we employ batch normalization, in which all responses are divided by the standard deviation taken over all spatial positions and all layers, and over a batch of input images (Ioffe & Szegedy, (2015)). The output stage of each layer in the network (following the subsampling stage) serve as the inputs to the Batch Normalization stage. For the outputs of a given layer for a given minibatch, $B = \{x1 \dots m\}$, the outputs of the corresponding Batch Normalization stage, $y_i = BN(x_i)$, is computed as follows: We first compute the minibatch mean, μ_B, and variance, σ_B^2, globally across all coefficients in

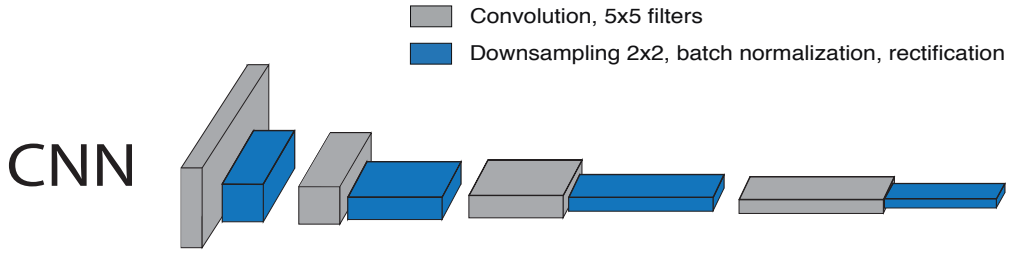

Figure 1.13: Architecture of a 4-layer Convolutional Neural Network (CNN). Each layer consists of a convolution, downsampling, and a rectifying nonlinearity (see text). The network was trained, using batch normalization, to maximize correlation with the TID-2008 database of human image distortion sensitivity.

B.

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1.1}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{1.2}$$

We then normalize the coefficients of B by μ_B and σ_B^2.

$$\hat{x}_i = \frac{(x_i - \mu_B)}{\sqrt{\sigma_B^2 + \epsilon}}; \tag{1.3}$$

We then scale and shift the output parameters, $\hat{x}_i$, with parameters that are learned from the data.

$$y_i = \lambda \hat{x}_i + \beta \tag{1.4}$$

Finally, outputs are rectified with a softplus nonlinearity, $\log(1 + \exp(x))$. After training, these normalization parameters (μ_B, σ_B^2) are fixed to the global mean and variance across the training set, and the scale and shift parameters (λ, and β), are fixed to the learned values.

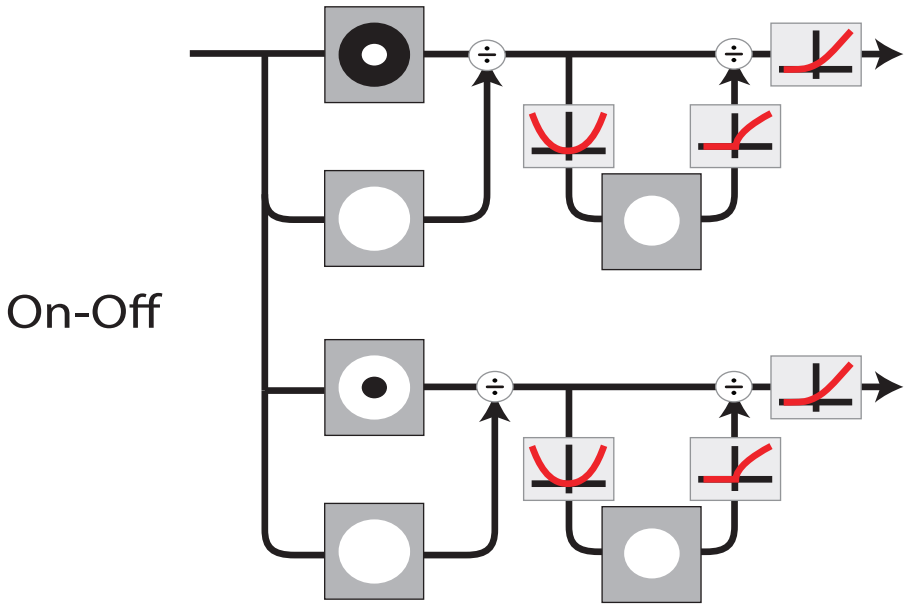

Figure 1.14: Architecture of our full LGN model (On-Off). (See text (and Appendix A) for details)

1.5.3 Models of Early Visual Physiology

We constructed a set of models reflecting the structure and computations of the Lateral Geniculate Nucleus (LGN), the visual relay center of the Thalamus. The full model (On-Off), is inspired by the model of Mante and Carandini, and is constructed from a cascade of linear filtering, and nonlinear computational modules (local gain control and rectification) (Mante et al., (2008)). The other three models are sub-models of the On-Off model.

On-Off Model

The first stage decomposes the image into two separate channels. Within each channel, the image is filtered by a difference-of-Gaussians (DoG) filter (2 parameters, controlling spatial size of the Gaussians - DoG filters in On and Off channels are assumed to be of

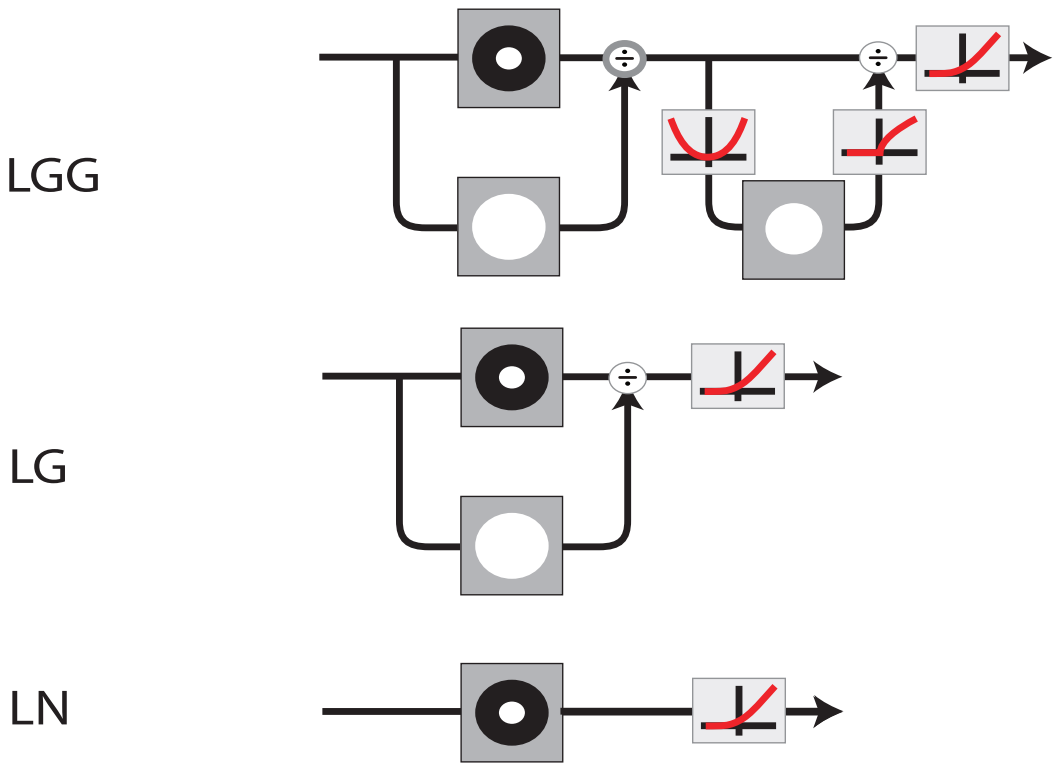

Figure 1.15: Architecture of our reduced LGN models. (See text (and Appendix A) for details)

opposite sign). Following this linear stage, the outputs are normalized by two sequential stages of gain control, a known property of LGN neurons (Mante et al. [2008]). Filter outputs are first normalized by a local measure of luminance (2 parameters, controlling filter size and amplitude), and subsequently by a local measure of contrast (2 parameters, again controlling size and amplitude). Finally, the outputs of each channel are rectified by a softplus nonlinearity, for a total of 12 model parameters.

In order to evaluate the necessity of each structural element of this model, we also test three reduced sub-models, each trained on the same data.

LGG Model

In the first reduced model (Linear-Gain Control-Gain Control, or LGG), we reduce the number of channels to one, but keep both stages of normalization. This model has half the parameters of the full On-Off model controlling the shape of the difference-of-gaussians filter, as well as the size of both normalization pools and the strength of each normalization (a total of 6 parameters).

LG Model

In the second reduced model (Linear-Gain Control, or LG), we reduce the number of channels to one, and also remove one stage of normalization. The parameters of this model control the shape of the difference-of-gaussians filter, as well as the size of the normalization pool and the strength of the normalization (a total of 4 parameters).

LN Model

In the final reduced model (Linear Nonlinear, or LN), we reduce the number of channels to one, and also remove both stages of normalization. The parameters of this model control the shape of the difference-of-gaussians filter (a total of 2 parameters).

1.6 Model Performance and Comparison with the State of the Art

After optimizing each of our networks on a subset of the TID-2008 database (900 distorted image pairs), we tested each model's performance at predicting a held-out set of testing data (800 distorted image pairs). We find that both deep neural networks, as well as the

TID 2008 Held-out Testing Set									
	PSNR	SSIM	V1	LN	LG	LGG	On-Off	VGG-IQA	CNN
ρ	0.52	0.74	0.81	0.66	0.74	0.83	0.82	0.84	0.86

Table 1.1: Evaluation of neural IQA models in the held-out testing set of TID2008 (Ponomarenko et al., 2009). Pearson correlation of distance metrics vs. human perceptual judgments. Numbers were obtained using the gray-scale version of the images in databases (see the text for details)

two LDN models that contain more than one stage of divisive normalization, perform at a high level on this held out dataset (See Table 1.1). In fact, we find that all four of these models outperform both PSNR, and two state-of-the-art metrics (SSIM and the V1 model of (Laparra et al., 2010). Unsurprisingly, our two simplest models (LN and LG) perform significantly worse.

1.7 New Applications Demand More of our Models

Surprisingly, we find that the most generic neural network, with $\sim 500,000$ parameters, and lacking any prior information about vision, or visual physiology, predicts our held out testing set slightly better than our more physiologically informed models. While this result may seem on immediate inspection like a victory for the generic neural network, the truly surprising result is that a model of the LGN with 12 parameters (the On-Off model) performs at nearly the same level as the much more complicated neural networks. The explanation for this lies partly in the particular construction of this database, which was created by image processing engineers, and represents the set of most commonly encountered image distortions within that community (Ponomarenko et al., (2009)). The space of possible image distortions, however, is very high-dimensional, and unlikely to be well spanned by this subset of distortions. While the models did not see the testing data during training, both subsets of data likely cover a similar subspace of the overarching distortion

space. While our models may generalize within, or near, this subspace, it is not clear that they will generalize correctly to distortions that lie in other parts of the space.

If our goal was only to build a model that could identify commonly encountered image distortions, and identify their visibility correctly, we may not worry so much about this problem. Recently, however, with the increase in computing power, and the advent of machine learning and high dimensional optimization, there has been a resurgence of interest in perceptual metrics. Many researchers are searching for perceptual loss functions that they can utilize to optimize auto-endcoders, end-to-end optimized image compression algorithms, generative adversarial networks, and many other algorithms. Utilized in this way, a good perceptual metric needs to generalize across the much larger space of possible distortions.

In the subsequent chapters of this thesis, we will explore the ability of each of our models to generalize beyond the database using a novel image synthesis test. After analyzing the performance of each of our models, we develop a novel image rendering framework, which requires a perceptual loss function, and analyze the performance of our models in this framework. Finally, we will make progress towards developing and testing a normative model of perceptual distortion sensitivity, based on the intuitions of Wang and Simoncelli that led them to develop SSIM (Wang et al., (2004)).

Chapter 2

Eliciting Predictions from High-Dimensional Representations

2.1 Eigen-distortions of Hierarchical Representations

Human capabilities for recognizing complex visual patterns are believed to arise through a cascade of transformations, implemented by neurons in successive stages in the visual system. Several recent studies have suggested that representations of deep convolutional neural networks trained for object recognition can predict activity in areas of the primate ventral visual stream better than models constructed explicitly for that purpose (Khaligh-Razavi & Kriegeskorte, (2014) and Yamins et al., (2014)). These results have inspired exploration of deep networks trained on object recognition as models of human perception, explicitly employing their representations as perceptual distortion metrics or loss functions (Dosovitskiy & Brox, (2016), Hénaff & Simoncelli, (2016), and Johnson et al., (2016)). On the other hand, several other studies have used synthesis techniques to generate images that indicate a profound mismatch between the sensitivity of these networks and that of human observers. Specifically, Szegedy et al., (2013) constructed image distortions, im-

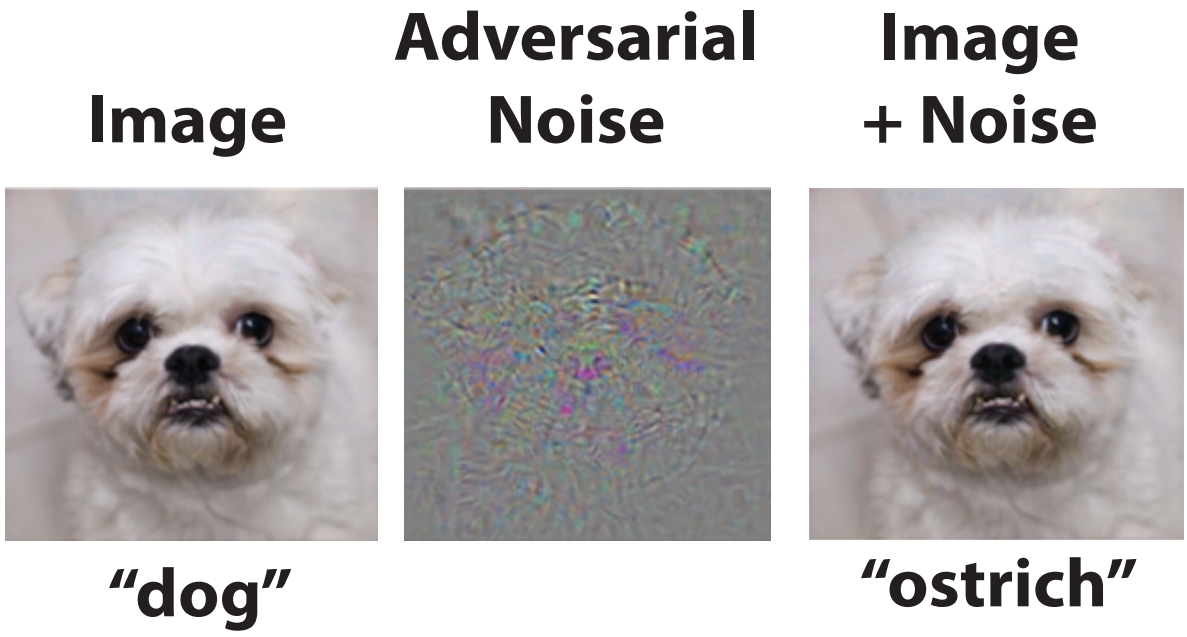

Figure 2.1: **Adversarial Examples** Szegedy et al. showed that classification performance within deep neural networks trained on object recognition is vulnerable to small targeted noise perturbations. In the example shown here, the neural network in question misclassfies the image of the dog on the right (which is a combination of the picture of the dog on the left, and the noise vector in the center) as an ostrich. Subsequent studies have found that different network architectures are vulnerable to the same image perturbations. This result indicates that neural networks and humans differ in their sensitivity to a least a subset of image distortions. (Adapted fromSzegedy et al., (2013))

perceptible to humans, that cause their networks to grossly misclassify objects (see figure 2.1). Similarly, Nguyen & Clune, (2015) optimized randomly initialized images to achieve reliable recognition by a network, but found that the resulting 'fooling images' were uninterpretable by human viewers (see figure 2.2). Simpler networks, designed for texture classification and constrained to mimic the early visual system, do not exhibit such failures (Portilla & Simoncelli, (2000)). These results have prompted efforts to understand why generalization failures of this type are so consistent across deep network architectures, and to develop more robust training methods to defend networks against attacks designed to exploit these weaknesses (Goodfellow et al., (2014)). From the perspective of modeling human perception, these synthesis failures suggest that representational spaces within deep neural networks deviate significantly from those of humans, and that methods for comparing representational similarity, based on fixed object classes and discrete sampling of the representational space, are insufficient to expose these deviations. If we are going to use such networks as models for human perception, we need better methods of comparing model representations to human vision. Recent work has taken the first step in this direction, by analyzing deep networks' robustness to visual distortions on classification tasks, as well as the similarity of classification errors that humans and deep networks make in the presence of the same kind of distortion (Dodge & Karam, (2017)).

Here, we aim to accomplish something in the same spirit, but rather than testing on a set of hand-selected examples, we develop a model-constrained synthesis method for generating targeted test stimuli that can be used to compare the layer-wise representational sensitivity of a model to human perceptual sensitivity. Synthesis tests, like that introduced by Freeman and Simoncelli (Freeman & Simoncelli, (2011)), give us the ability to compare model representations with human representations. In their paper, Freeman and Simoncelli

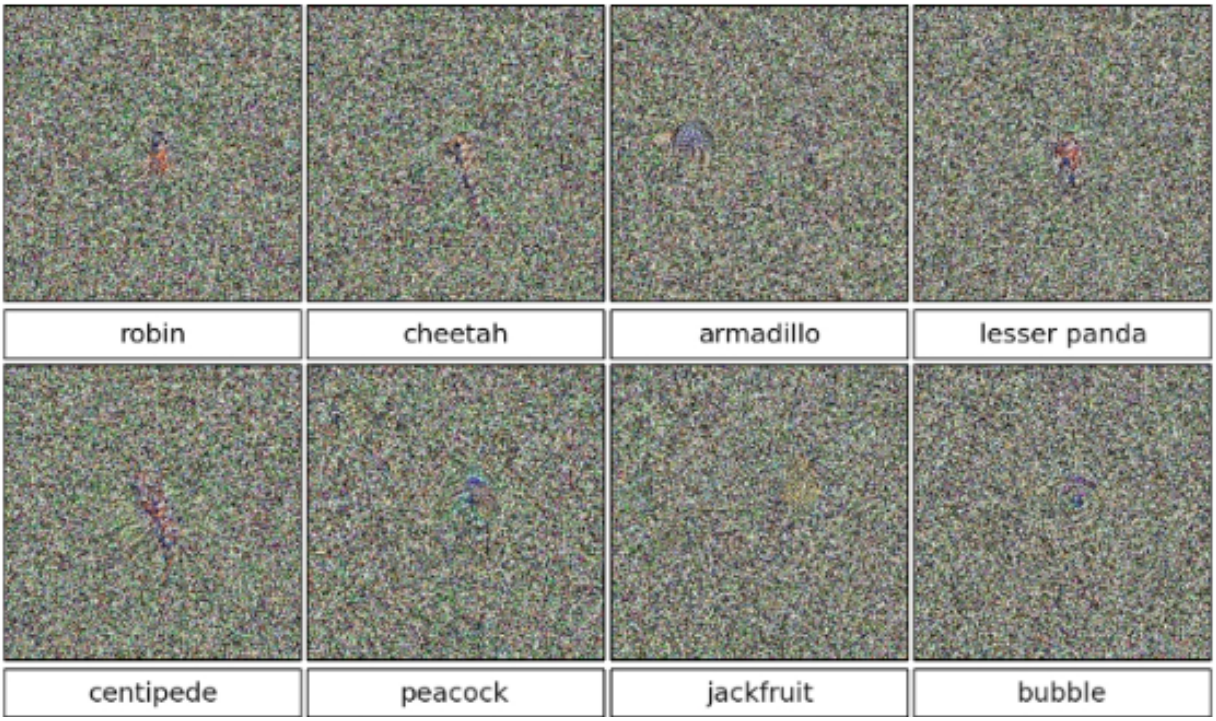

Figure 2.2: **Fooling Images**: The above images were synthesized from white noise samples to maximize network confidence that the object listed below the image was present. The authors dubbed these images, "Fooling Images" because the neural network is fooled, despite the fact that humans can tell clearly that the object in question is not present in the image. This result suggests that the network may be overly tolerant to unnatural image features. (Adapted from Nguyen & Clune, (2015))

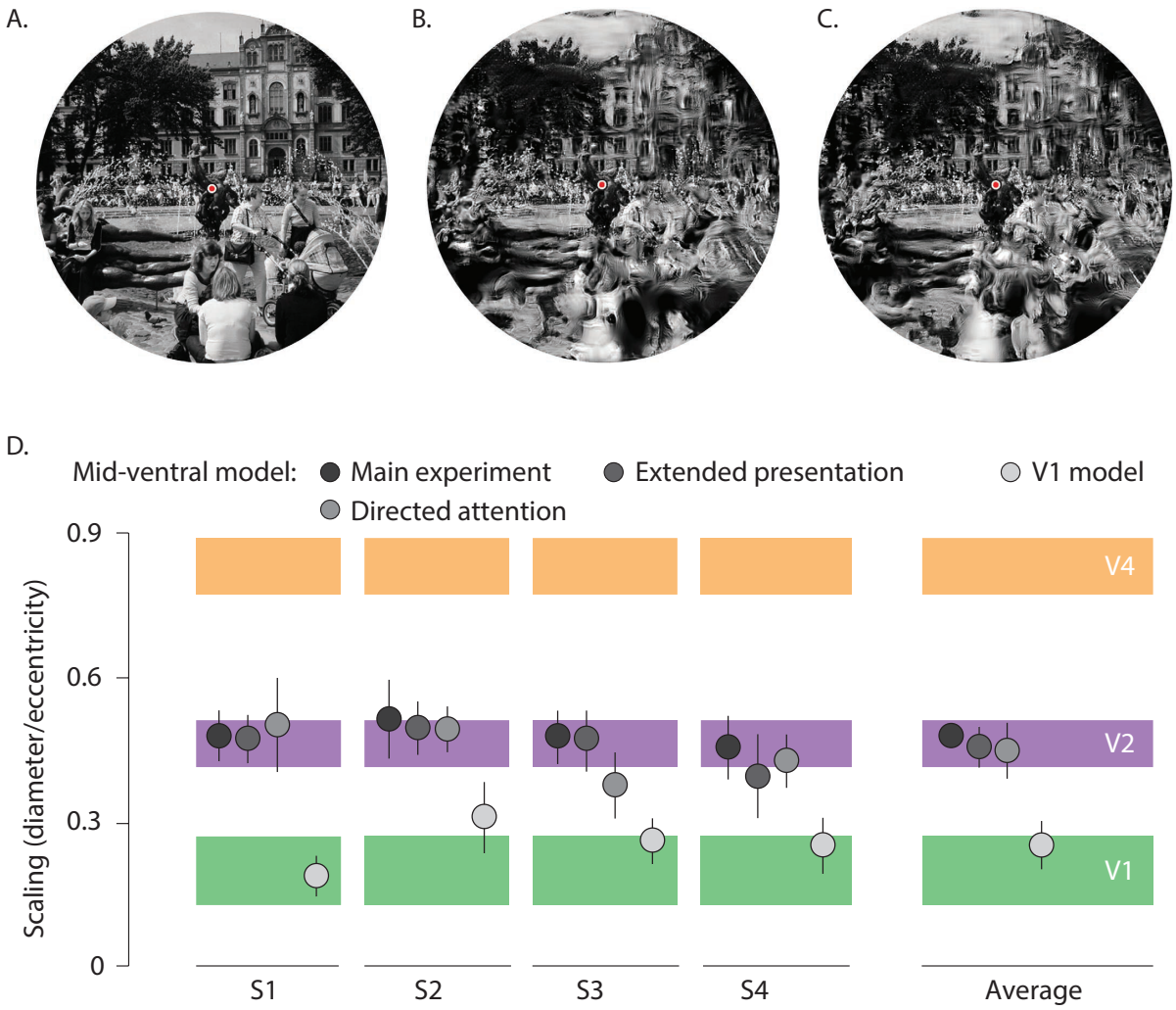

Figure 2.3: **Metamers of the ventral stream** Freeman and Simoncelli introduced a method for generating multiple images which differed only along dimensions within the null space of a model (in their case, a model of the intermediate layers of the ventral visual stream) (Freeman & Simoncelli, (2011)). They refer to these images as metamers. Images B. and C. are model synthesized metamers of the original image, A. (D.) The authors showed that when the size of the receptive fields in the model used to generate the images matched the size of receptive fields in area V2, the human observers were unable to distinguish between metameric images. (Adapted from Freeman & Simoncelli, (2011))

introduced a method for generating multiple images which differed only along dimensions within the null space of a model (in their case, a model of the intermediate layers of the ventral visual stream). They referred to these images as ventral-stream metamers, and were able to show that humans cannot distinguish between two different, but metameric, images when the size of the receptive fields in the model used to generate the images matched the size of receptive fields in area V2 (see figure 2.3). This method is a strong test of the similarity of the model representation and the human perceptual representation.

Similarly, we could synthesize images that differ along dimensions that lie outside of the null space of the model. Unlike metamers, these differences between images should be noticeable to human observers. Not every difference will be equally noticeable, however, and the model's predictions of the ranking of detectability of these differences is also a strong test of the similarity of a model representation and the human perceptual representation.

Utilizing Fisher information, we isolate the model-predicted most and least noticeable changes to an image. We test these predictions by determining how well human observers can discriminate these same changes (see figure 2.4). We apply this method to six layers of VGG16 (Simonyan & Zisserman, (2015)), a deep convolutional neural network (CNN) trained to classify objects. We also apply the method to several models explicitly trained to predict human sensitivity to image distortions, including both a 4-stage generic CNN, an optimally-weighted version of VGG16, and a family of highly-structured models explicitly constructed to mimic the physiology of the early human visual system. Example images from the paper, as well as additional examples, are available at `http://www.cns.nyu.edu/~lcv/eigendistortions/`.

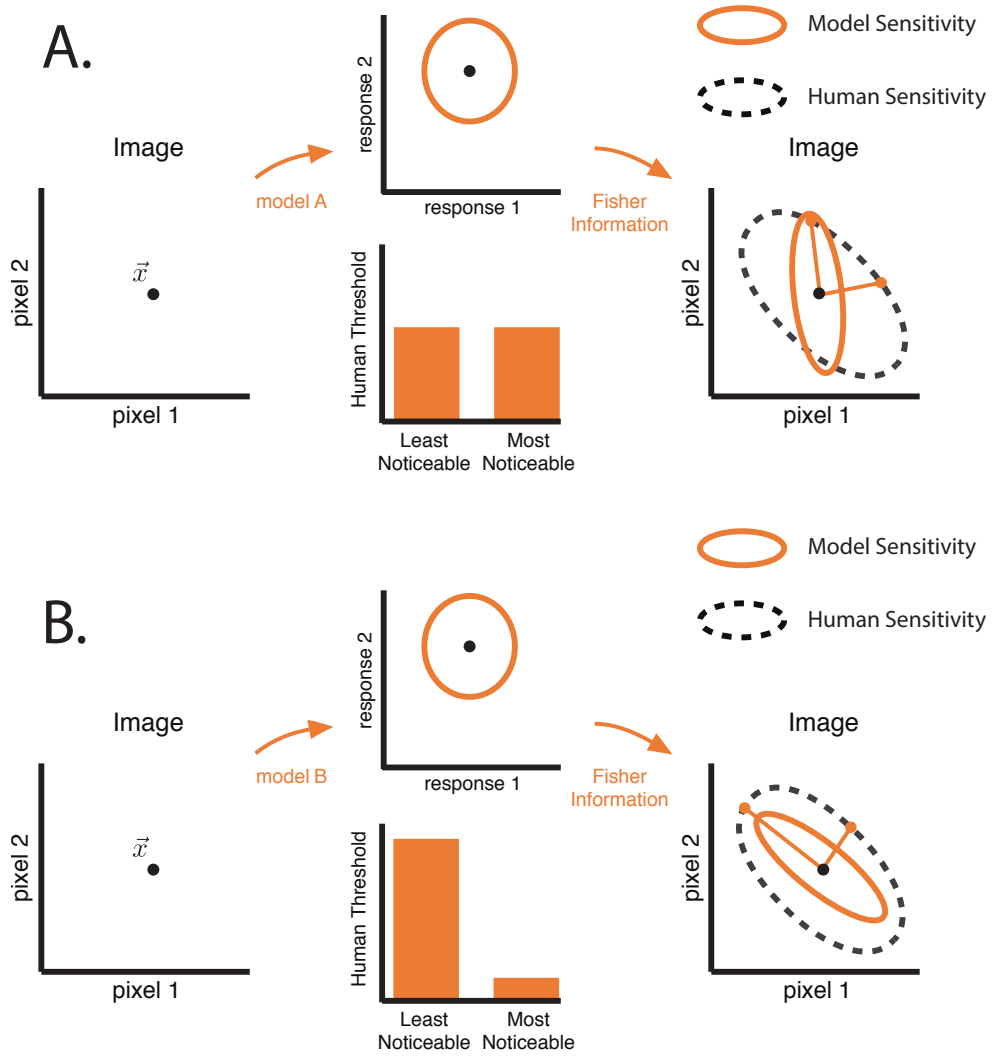

Figure 2.4: **Using Fisher Information to elicit extremal predictions from each model**:A model is are applied to an image (depicted as a point $\vec{x}$ in the space of pixel values), producing a response vector. The level set of equivalently detectable image distortions lie on a circle (or hypersphere in higher dimensions) around the response to this image, because we utilize MSE to measure distance in this space. The Fisher Information Matrix (FIM) of the model allows us to describe this level set in the pixel space. We test the model's predictions of the most and least-noticeable distortions by measuring human discriminability in these directions. If the model is a bad model for human sensitivity, the detection thresholds in the directions of these two extremal predictions will be similar (A.). If the model is a good model for human sensitivity, the detection threshold in the direction of the least noticeable prediction of the model will be higher than the detection threshold for the least noticeable prediction (B.).

2.2 Synthesizing Model Predictions

2.2.1 Predicting Discrimination Thresholds

Suppose we have a model for human visual representation, defined by conditional density $p(\vec{r}|\vec{x})$, where $\vec{x}$ is an N-dimensional vector containing the image pixels, and $\vec{r}$ is an M-dimensional random vector representing responses internal to the visual system (e.g., firing rates of a population of neurons). For our analysis, we require If the image is modified by the addition of a distortion vector, $\vec{x} + \alpha\hat{u}$, where $\hat{u}$ is a unit vector, and scalar α controls the amplitude of distortion, the model can be used to predict the threshold at which the distorted image can be reliably distinguished from the original image. Specifically, one can express a lower bound on the discrimination threshold in direction $\hat{u}$ for any observer or model that bases its judgments on $\vec{r}$ (Seriès et al., (2009)):

$$T(\hat{u};\vec{x}) \geq \beta\sqrt{\hat{u}^T J^{-1}[\vec{x}]\hat{u}} \tag{2.1}$$

where β is a scale factor that depends on the noise amplitude of the internal representation (as well as experimental conditions, when measuring discrimination thresholds of human observers), and $J[\vec{x}]$ is the Fisher information matrix (FIM; Fisher, (1925)), a first-order expansion of the log likelihood:

$$J[\vec{x}] = \mathbb{E}_{\vec{r}|\vec{x}}\left[\left(\frac{\partial}{\partial\vec{x}}\log p(\vec{r}|\vec{x})\right)\left(\frac{\partial}{\partial\vec{x}}\log p(\vec{r}|\vec{x})\right)^T\right] \tag{2.2}$$

Here, we restrict ourselves to models that can be expressed as a deterministic (and differentiable) mapping from the input pixels to mean output response vector, $f(\vec{x})$, with additive white Gaussian noise in the response space. The log likelihood in this case reduces to a

quadratic form:

$$\log p(\vec{r}|\vec{x}) = -\frac{1}{2}\Big([\vec{r} - f(\vec{x})]^T[\vec{r} - f(\vec{x})]\Big) + \text{const.}$$

The derivative of the $\log p(\vec{r}|\vec{x})$ with respect to x is:

$$\frac{\partial}{\partial \vec{x}} \log p(\vec{r}|\vec{x}) = \frac{\partial f}{\partial \vec{x}}^T [r - f(x)]$$

Substituting this into Eq. (2.2) gives:

$$J[\vec{x}] = \mathbb{E}_{\vec{r}|\vec{x}}\left[\frac{\partial f}{\partial \vec{x}}^T [r - f(x)][r - f(x)]^T \frac{\partial f}{\partial \vec{x}}\right]$$

The expectation of $[r - f(x)][r - f(x)]^T = \Sigma = I$, and can be dropped, giving:

$$J[\vec{x}] = \frac{\partial f}{\partial \vec{x}}^T \frac{\partial f}{\partial \vec{x}}$$

Thus, for these models, the Fisher information matrix induces a locally adaptive Euclidean metric on the space of images, as specified by the Jacobian matrix, $\partial f/\partial \vec{x}$. Our analysis relies on an invertible $J[\vec{x}]$, which requires that $M \geq N$, or that the representation space is either complete or overcomplete. An undercomplete representation results in an $J[\vec{x}]$ that is singular, and is not invertible. Thus, for all models explored in this section, we require that $M \geq N$.

2.2.2 Extremal Eigen-Distortions

The FIM is generally too large to be stored in memory or inverted. Even if we could store and invert it, the high dimensionality of input (pixel) space renders the set of possible distortions too large to test experimentally. We resolve both of these issues by restricting

our consideration to the most- and least-noticeable distortion directions, corresponding to the eigenvectors of $J[\vec{x}]$ with largest and smallest eigenvalues, respectively. First, note that if a distortion direction $\hat{e}$ is an eigenvector of $J[\vec{x}]$ with associated eigenvalue λ, then it is also an eigenvector of $J^{-1}[\vec{x}]$ (with eigenvalue $1/\lambda$), since the FIM is symmetric and positive semi-definite. In this case, Eq. (2.1) becomes

$$T(\hat{e};\vec{x}) \geq \beta/\sqrt{\lambda}$$

If human discrimination thresholds attain this bound, or are a constant multiple above it, then the ratio of discrimination thresholds along two different eigenvectors is the square root of the ratio of their associated eigenvalues. In this case, the strongest prediction arising from a given model is the ratio of the *extremal* (maximal and minimal) eigenvalues of its FIM, which can be compared to the ratio of human discrimination thresholds for distortions in the directions of the corresponding extremal eigenvectors (Fig. 2.5).

Although the FIM cannot be stored, it is straightforward to compute its product with an input vector (i.e., an image). We can decompose this product into a series of Matrix-vector products, requiring us to store only a vector the size of the input image at each stage. Using this operation, we can solve for the extremal eigenvectors using the well-known power iteration method (Mises & Pollaczek-Geiringer, (1929)). Specifically, to obtain the maximal eigenvalue of a given function and its associated eigenvector (λ_m and $\hat{e}_m$, respectively), we start with a vector consisting of white noise, $\hat{e}_m^{(0)}$, and then iteratively apply the FIM, renormalizing the resulting vector, until convergence:

$$\lambda_m^{(k+1)} = \left\| J[\vec{x}]\hat{e}_m^{(k)} \right\|; \quad \hat{e}_m^{(k+1)} = J[\vec{x}]\hat{e}_m^{(k)}/\lambda_m^{(k+1)}$$

To obtain the minimal eigenvector, $\hat{e}_l$, we perform a second iteration using the FIM with

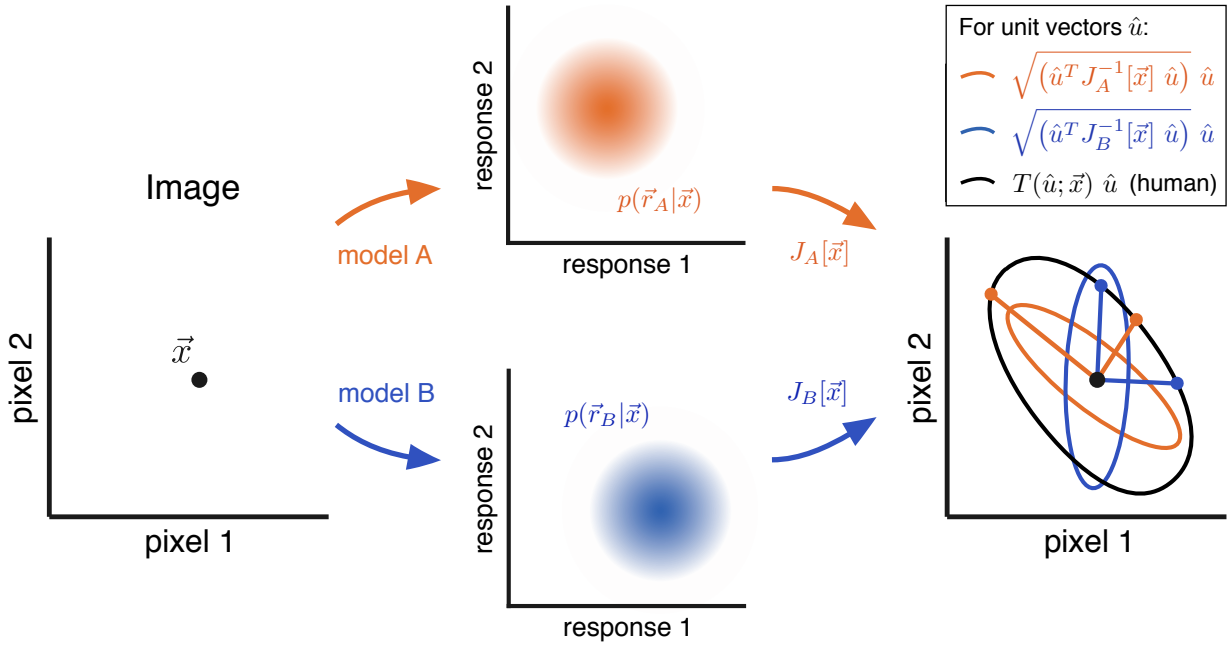

Figure 2.5: Two models are applied to an image (depicted as a point $\vec{x}$ in the space of pixel values), producing response vectors $\vec{r}_A$ and $\vec{r}_B$. Responses are assumed to be stochastic, and drawn from known distributions $p(\vec{r}_A|\vec{x})$ and $p(\vec{r}_B|\vec{x})$. The Fisher Information Matrices (FIM) of the models, $J_A[\vec{x}]$ and $J_B[\vec{x}]$, provide a quadratic approximation of the discriminability of distortions relative to an image (rightmost plot, colored ellipses). The extremal eigenvalues and eigenvectors of the FIMs (directions indicated by colored lines) provide predictions of the most and least visible distortions. We test these predictions by measuring human discriminability in these directions (colored points). In this example, the ratio of discriminability along the extremal eigenvectors is larger for model A than for model B, indicating that model A provides a better description of human perception of distortions (for this image).

the maximal eigenvalue subtracted from the diagonal:

$$\lambda_l{}^{(k+1)} = \left\| (J[\vec{x}] - \lambda_m I)\, \hat{e}_l{}^{(k)} \right\|; \quad \hat{e}_l{}^{(k+1)} = (J[\vec{x}] - \lambda_m I)\, \hat{e}_l{}^{(k)} / \lambda_l^{(k+1)}$$

2.2.3 Measuring Human Detection Thresholds

For each model under consideration, we synthesized extremal eigen-distortions for 6 images from the Kodak image set[1]. We then estimated human thresholds for detecting these distortions using a two-alternative forced-choice task. On each trial, subjects were shown (for one second each with a half second blank screen between images, and in randomized order) a photographic image (18 degrees across), $\vec{x}$, and the same image distorted using one of the extremal eigenvectors, $\vec{x}+\alpha\hat{e}$, and then asked to indicate which image appeared more distorted. This procedure was repeated for 120 trials for each distortion vector, $\hat{e}$, over a range of α values, with ordering chosen by a standard psychophysical staircase procedure. The proportion of correct responses, as a function of α, was fit with a cumulative Gaussian function (see Appendix B), and the subject's detection threshold, $T_s(\hat{e};\vec{x})$ was estimated as the value of α for which the subject could distinguish the distorted image 75% of the time. We computed the natural logarithm of the ratio of these detection thresholds for the minimal and maximal eigenvectors, and averaged this over images (indexed by i) and subjects (indexed by s):

$$D(f) = \frac{1}{S}\frac{1}{I}\sum_{s=1}^{S}\sum_{i=1}^{I} \log \left\| T_s(\hat{e}_{li};\vec{x}_i)/T_s(\hat{e}_{mi};\vec{x}_i) \right\|$$

where T_s indicates the threshold measured for human subject s. $D(f)$ provides a measure of a model's ability to predict human performance with respect to distortion detection: the

[1]Downloaded from `http://www.cipr.rpi.edu/resource/stills/kodak.html`.

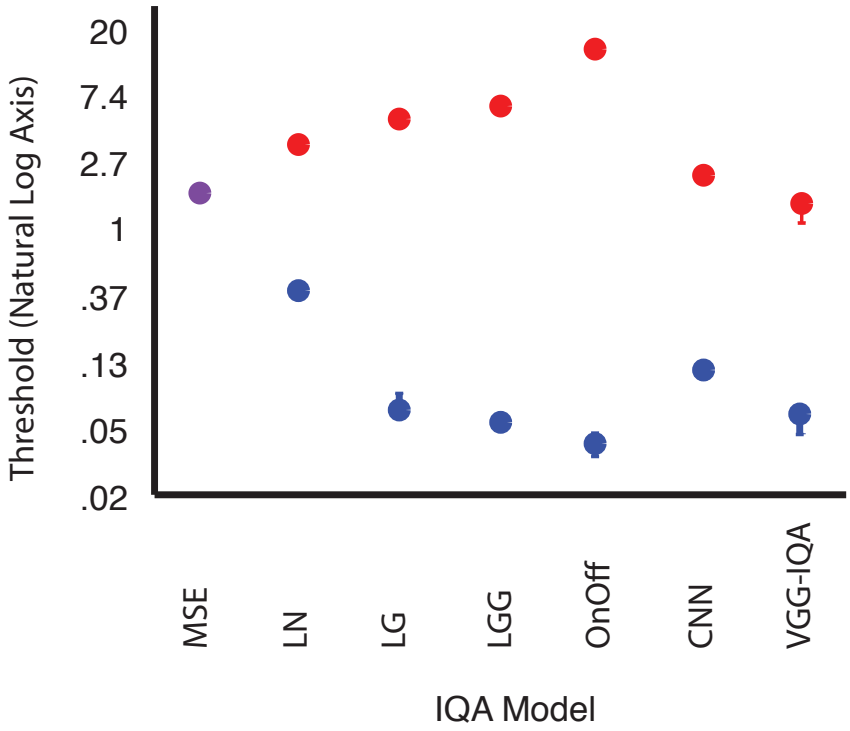

Figure 2.6: **Top:** Average log-thresholds for detection of the least-noticeable (red) and most-noticeable (blue) eigen-distortions derived from IQA models (19 human observers).

ratio of thresholds for model-generated extremal distortions will be larger for models that are more similar to the human subjects (Fig. 2.5).

2.3 Quantifying the Quality of Model Predictions

2.3.1 Comparing Perceptual Predictions of Generic and Structured Models

After training, we evaluated each model's predictive performance using traditional cross-validation methods on a held-out test set of the TID-2008 database. By this measure, all three models performed well (Pearson correlation: CNN $\rho = .86$, On-Off: $\rho = .82$, VGG-IQA: $\rho = .84$).

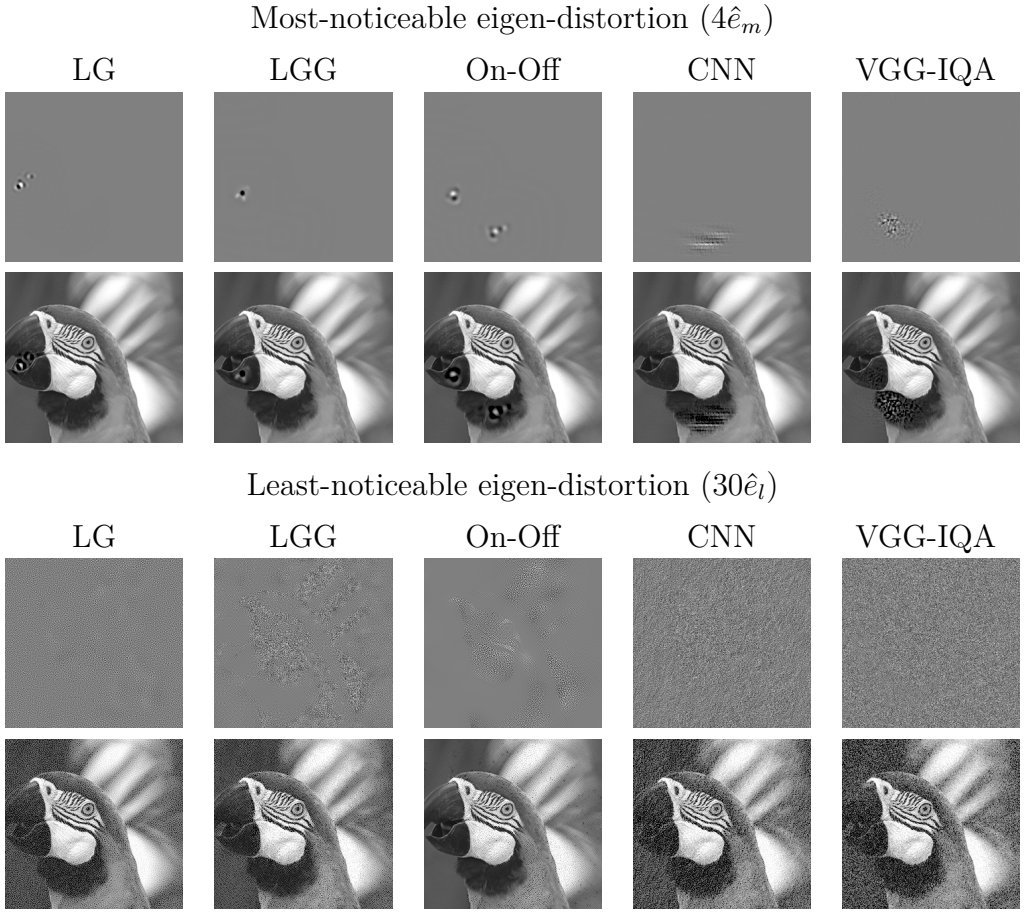

Figure 2.7: Eigen-distortions for several models trained to maximize correlation with human distortion ratings in TID-2008 Ponomarenko et al., (2009). Images are best viewed in a display with luminance range from 5 to 300 cd/m^2 and a γ exponent of 2.4. **Top**: Most-noticeable eigen-distortions. All distortion image intensities are re-scaled by the same amount ($\times 4$). **Second row**: Original image ($\vec{x}$), and sum of this image with each eigen-distortion. **Third and fourth rows**: Same, for the least-noticeable eigen-distortions. All distortion image intensities re-scaled by the same amount ($\times 30$).

Stepping beyond the TID-2008 database, and using the more stringent eigen-distortion test, yielded a very different outcome (Figs. 2.7, 2.6 and 2.10). The average detection thresholds measured across 19 human subjects and 6 base images indicates that all of our models surpassed the baseline model in at least one of their predictions. However, the eigen-distortions derived from the generic CNN and VGG-IQA were significantly less predictive of human sensitivity than those derived from the On-Off model (Fig. 2.6) and, surprisingly, even somewhat less predictive than early layers of VGG16 (see Fig. 2.10). Thus, the eigen-distortion test reveals generalization failures in the CNN and VGG16 architectures that are not exposed by traditional methods of cross-validation. On the other hand, the models with architectures that mimic biology (On-Off, LGG, LG) are constrained in a way that enables better generalization.

We compared these results to the performance of each of our reduced LGN models (Fig. 1.15), to determine the necessity of each structural element of the full On-Off model. As expected, the models incorporating more LGN functional elements performed better on a traditional cross-validation test, with the most complex of the reduced models (LGG) performing at the same level as On-Off and the CNN (LN: $\rho = .66$, LG: $\rho = .74$, LGG: $\rho = .83$). Likewise, models with more LGN functional elements produced eigen-distortions with increasing predictive accuracy (Fig. 2.6 and 2.10). It is worth noting that the three LGN models that incorporate some form of local gain control perform significantly better than the CNN and VGG-IQA models, and better than all layers of VGG16, including the early layers (see Fig. 2.10).

2.3.2 Probing Representational Sensitivity of VGG16 Layers

We also examined discrimination predictions derived from several layers of original VGG16 model, which has been previously studied in the context of perceptual sensitivity. Specifically, Johnson et al., (2016) trained a neural network to generate super-resolution images using the representation of an intermediate layer of VGG16 as a perceptual loss function, and showed that the images this network produced looked significantly better than images generated with simpler loss functions (e.g. pixel-domain mean squared error). Hénaff & Simoncelli, (2016) used VGG16 as an image metric to synthesize minimal length paths (geodesics) between images modified by simple global transformations (rotation, dilation, etc.). The authors found that a modified version of the network produced geodesics that captured these global transformations well (as measured perceptually), especially in deeper layers. Implicit in both of these studies, and others like them (e.g., Dosovitskiy & Brox, (2016)), is the idea that a deep neural network trained to recognize objects may exhibit additional human perceptual characteristics.

Here, we compare VGG16's sensitivity to distortions directly to human perceptual sensitivity to the same distortions. We transformed luminance-valued images and distortion vectors to proper inputs for VGG16 following the preprocessing steps described in the original paper, and verified that our implementation replicated the published object recognition results. For human perceptual measurements, all images were transformed to produce the same luminance values on our calibrated display as those assumed by the model.

We computed eigen-distortions of VGG16 at 6 different layers: the rectified convolutional layer immediately prior to the first max-pooling operation (Front), as well as each subsequent layer following a pooling operation (Layer2–Layer6). A subset of the eigen-distortions are shown, both in isolation and superimposed on the image from which they

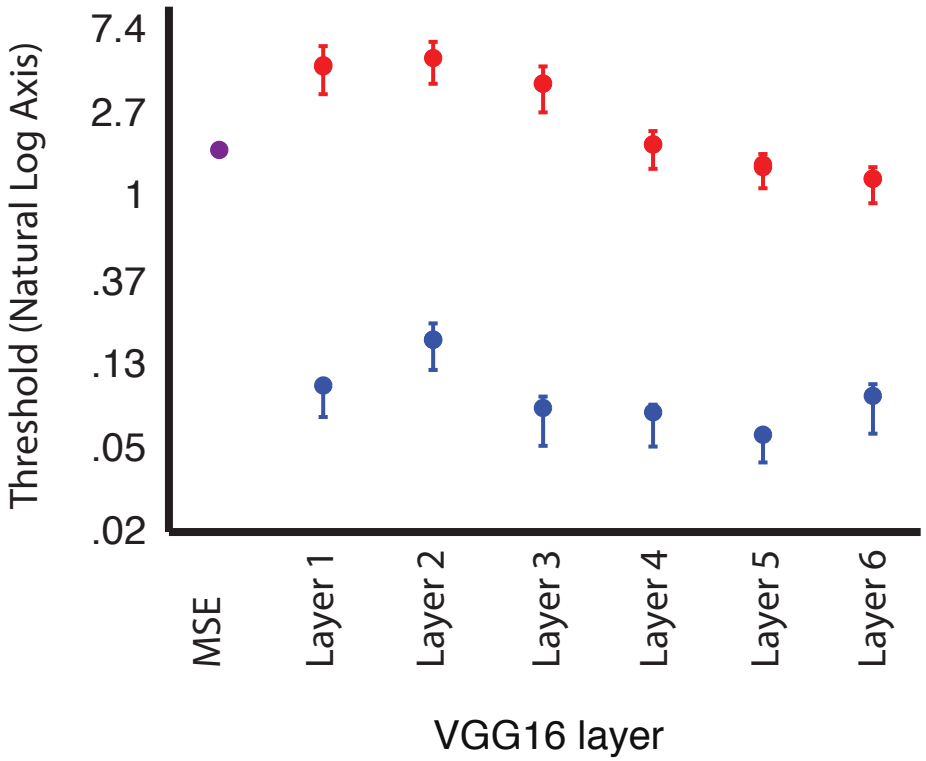

Figure 2.8: **Top:** Average log-thresholds for detection of the least-noticeable (red) and most-noticeable (blue) eigen-distortions derived from layers within VGG16 (10 observers), and a baseline model (MSE) for which distortions in all directions are equally visible.

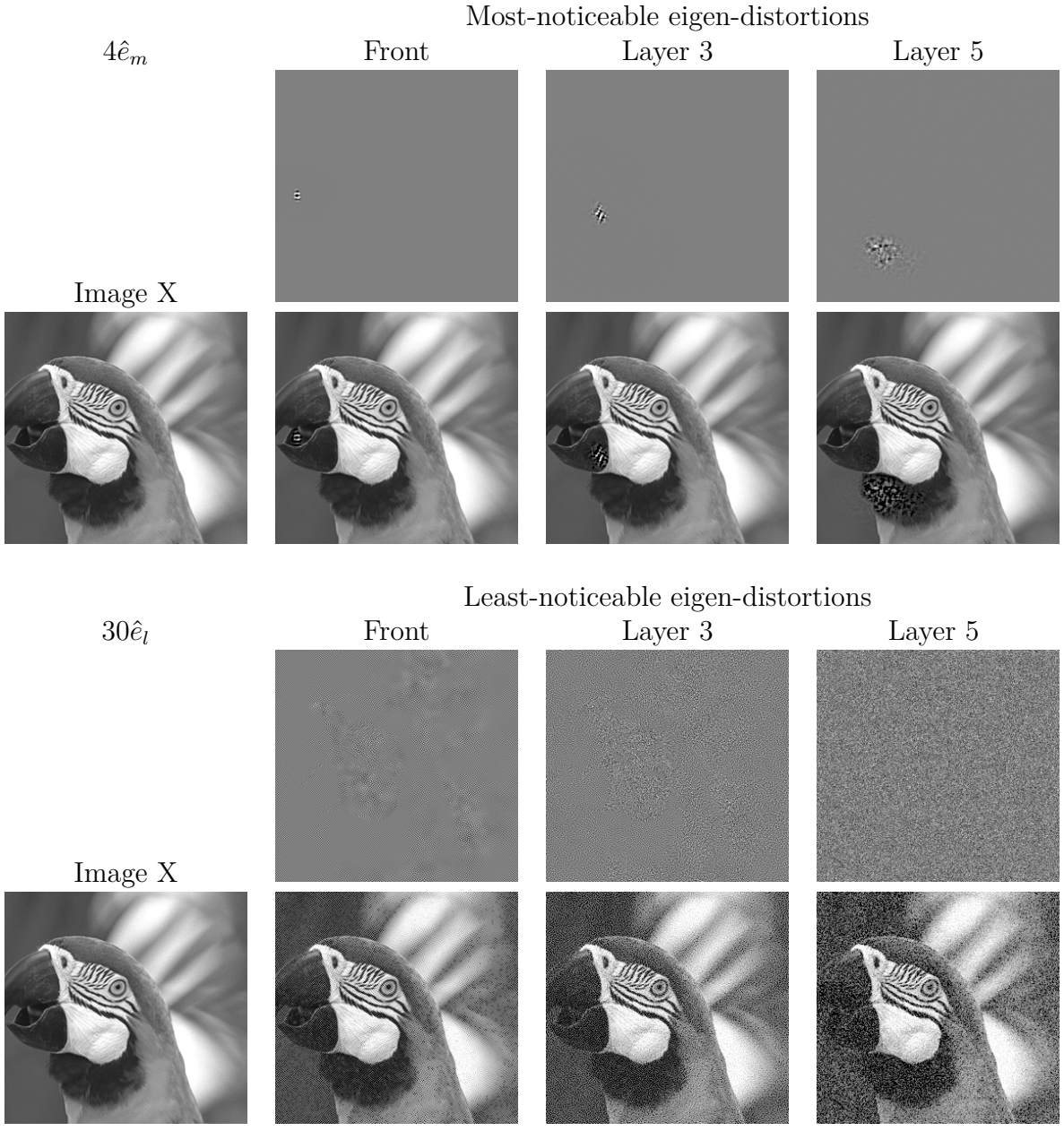

Figure 2.9: Eigen-distortions derived from three layers of the VGG16 network for an example image. Images are best viewed in a display with luminance range from 5 to 300 cd/m^2 and a γ exponent of 2.4. **Top**: Most-noticeable eigen-distortions. All distortion image intensities are scaled by the same amount ($\times 4$). **Second row**: Original image ($\vec{x}$), and sum of this image with each of the eigen-distortions. **Third and fourth rows**: Same, for the least-noticeable eigen-distortions. Distortion image intensities are scaled the same ($\times 30$).

were derived, in Fig. 2.9. Average Human detection thresholds measured across 10 subjects and 6 base images are summarized in Fig. 2.8.Note that the detectability of these distortions in isolation is not necessarily indicative of their detectability when superimposed on the underlying image, as measured in our experiments. We compared all of these predictions to a baseline model (MSE), where the image transformation, $f(\vec{x})$, is replaced by the identity matrix. For this model, every distortion direction is equally discriminable, and distortions are generated as samples of Gaussian white noise.

The results from our eigen-distortion analysis indicate that the early layers of VGG16 (in particular, Front and Layer3) are better predictors of extremal human sensitivity than the deeper layers (Layer4, Layer5, Layer6). Specifically, the most noticeable eigen-distortions from representations within VGG16 become more discriminable with depth, but so generally do the least-noticeable eigen-distortions. This discrepancy could arise from overlearned invariances, or invariances induced by network architecture (e.g. layer 6, the first stage in the network where the number of output coefficients falls below the number of input pixels, is an under-complete representation). Notably, including the "L2 pooling" modification suggested in Hénaff & Simoncelli, (2016) did not significantly alter the visibility of eigen-distortions synthesized from VGG16 (images and data not shown).

2.3.3 Comparing Model Sensitivity Predictions to Human Sensitivity

In the above section, we compared only whether each model's predictions of most- and least-noticeable distortion directions (the eigenvectors of their Fisher Information matrix) was well aligned with human perceptual sensitivity. The eigenvalues associated with those eigenvectors also carry a testable prediction. Specifically, the model's prediction of human detection threshold in the direction of the eigenvectors is proportional to the square root

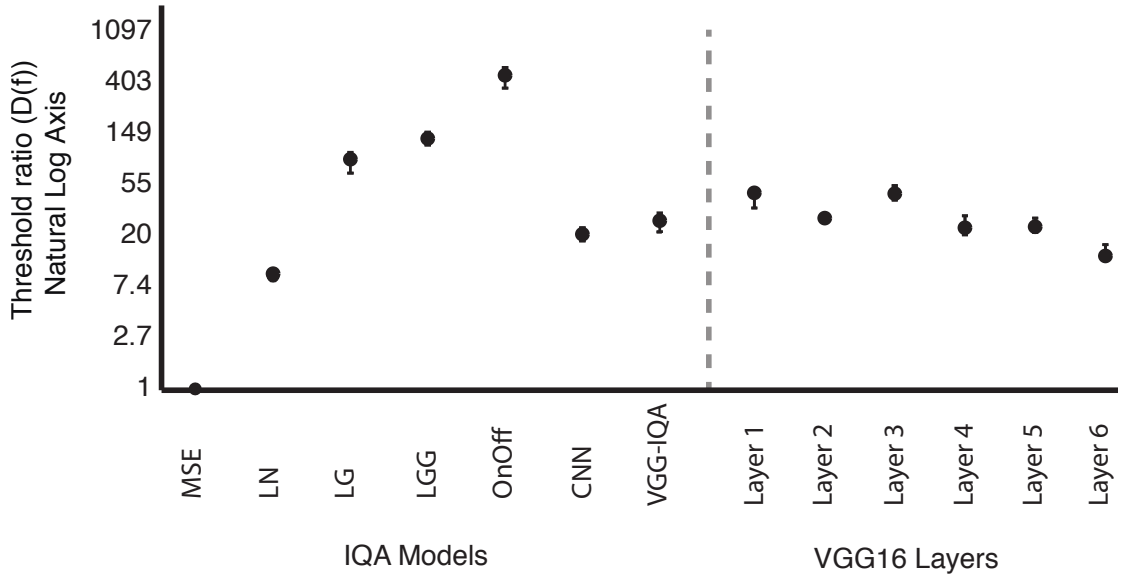

Figure 2.10: Average empirical log-threshold ratio (D) for eigen-distortions derived from each IQA optimized model and each layer of VGG16.

of the associated eigenvalue.

$$T(\hat{e}; \vec{x}) \geq \beta/\sqrt{\lambda}$$

If human discrimination thresholds attain this bound, or are a constant multiple above it, then the ratio of discrimination thresholds along two different eigenvectors is the square root of the ratio of their associated eigenvalues. In this case, the strongest prediction arising from a given model is the ratio of the *extremal* (maximal and minimal) eigenvalues of its FIM, which can be compared to the ratio of human discrimination thresholds for distortions in the directions of the corresponding extremal eigenvectors (Fig. 2.10).

We can test the quality of each model's predictions by comparing the square root of the ratio of its eigen-distortion eignvalues, and compare them to the ratio of measured human detection thresholds for the same images (D) (see Figure 2.11). This measure gives us a more nuanced picture of the ability of a model to capture human sensitivity. Both this

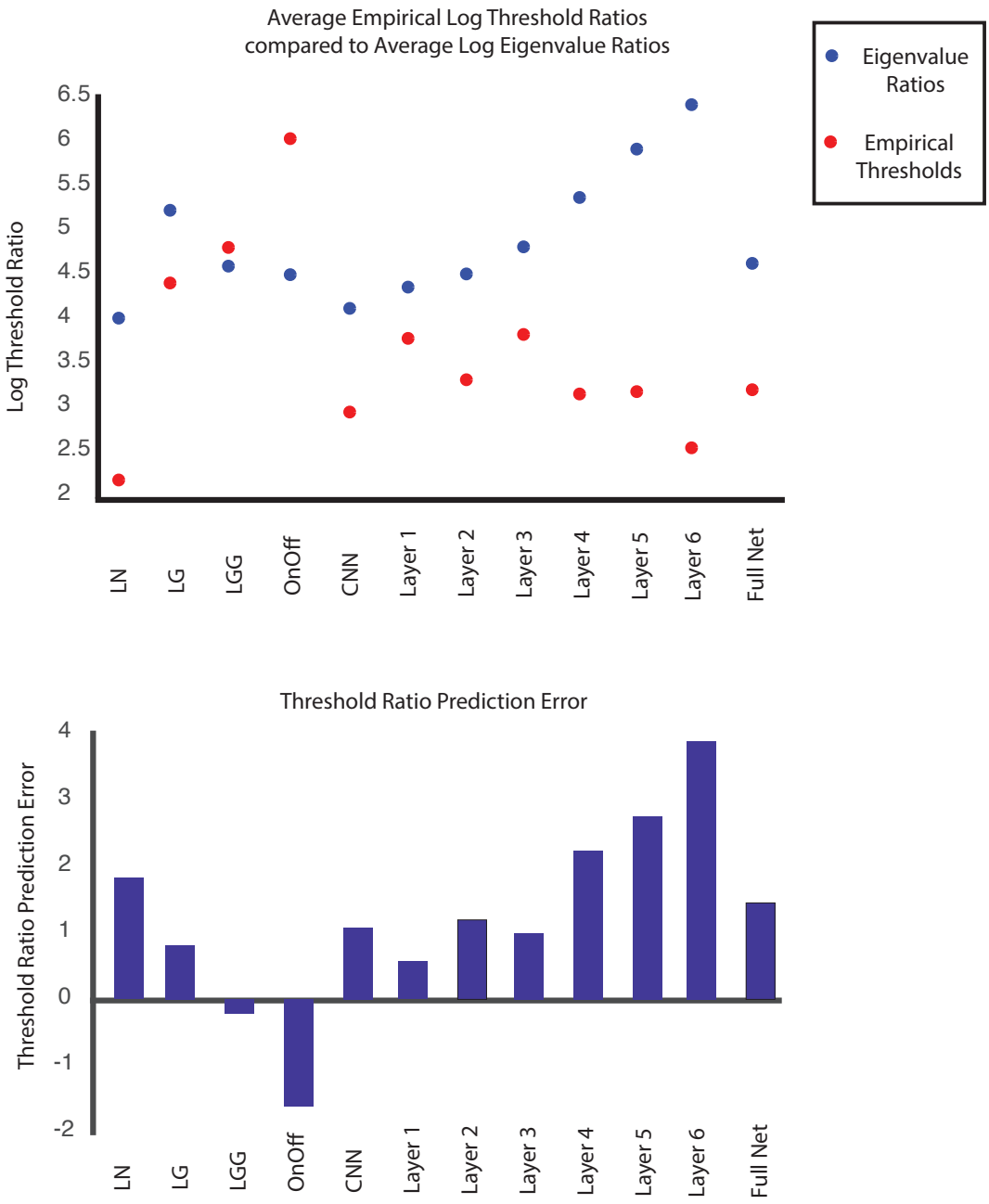

Figure 2.11: Comparison of Eigenvalue Ratios to Empirical Threshold Ratios for Each Model's Eigen-distortions shows that no models perfectly predict human sensitivity ratios. The models that under predict their own eigen-distortion detection ratio (LGG and OnOff) also found the set of distortions with the largest empirical ratios. At the same time, deeper layers of VGG produce larger and larger predicted threshold ratios while the empirically measured ratios get smaller and smaller.

measure, and the measures reported above, must be taken in tandem to understand the quality of a model.

An analysis of the models tested here suggests that none of our models perfectly predict the empirical threshold ratio of their own eigen-distortions, with LGG as the best performing model (See Figure 2.11). This is not entirely surprising, and it is important to take these results together with the ranking of the empirical detection threshold ratios across models. In our data, the models that under predict their own eigen-distortion detection ratio (LGG and OnOff) also found the set of distortions with the largest empirical ratios. At the same time, deeper layers of VGG produce larger and larger predicted threshold ratios while the empirically measured ratios get smaller and smaller.

2.3.4 Models as Observers

In addition to comparing model predictions about the detection ratios of their own eigen-distortions, we would like to quantify how well each model predicts the detectability of every other model's eigen-distortions. This analysis is partly related to the above analysis (models that over predict their own eigen-distortion detection ratio are incapable of accounting for the eigen-distortions of other model's with larger detection thresholds than their own).

To compare model derived distances for the eigendistortions across models, we can calculate the distance between our original image, and each distorted image shown to a subject in each of the model's response spaces (See Figure 2.12). We first show each model every eigen-distortion scaled at the same amplitude $\alpha = .1$. We then find the Pearson correlation between the measured distances from each model and the mean detection thresholds for each Eigen-distortion (See Figure 2.13.

Distances measured from the three models containing gain control modules (LG, LGG,

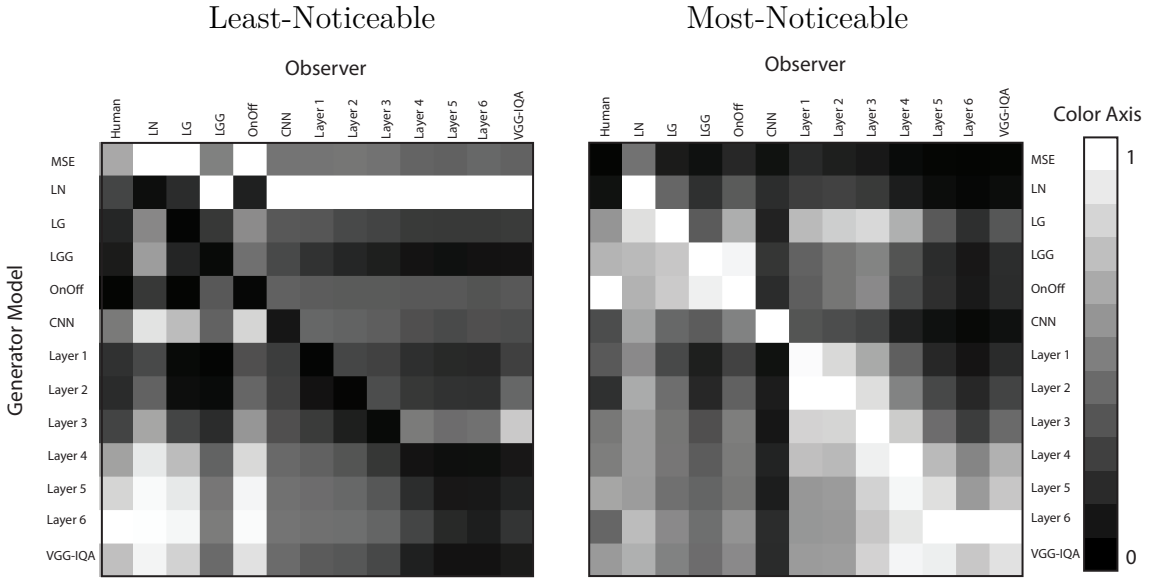

Figure 2.12: Observer (columns) Distances for Every Eigen-Distortion scaled at a single amplitude $\alpha = .1$ derived from each generator model (rows) (distances are normalized within each column for display purposes)

and OnOff) have the highest correlation with empirical human detection thresholds. The model with the next highest correlation, VGG layer 3, performs significantly worse than any of these modelss

The results above show that for small distortions, distances measured within models that contain gain control (especially within our OnOff model) best explain the observed human detection thresholds. We can also ask how well this result holds for over larger distortion amplitudes. To do so, we show each model each eigen-distortion scaled at every amplitude that our human subjects saw during the experiment. Taking all of this data together, we ask how likely the measured model distances at each of these amplitudes explains the observed hits and misses from our psychophysical measurements (see Appendix B for details and Figure 2.13 for data). The results in Figure 2.13 show that, over the range of amplitudes shown to human observers in our experiments, the OnOff model best explains the observed

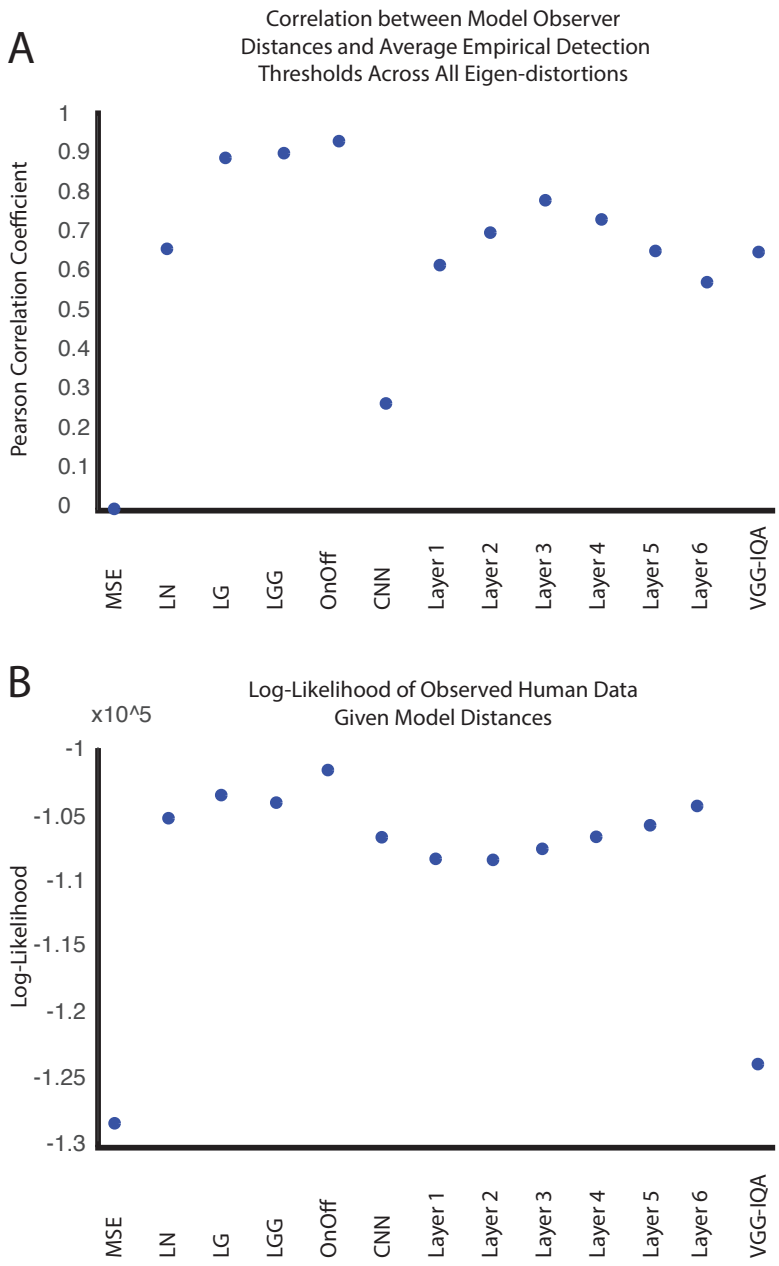

Figure 2.13: (A.) Pearson Correlation between Model Observer Distances and Average Empirical Detection Thresholds Across All Eigen-distortions. Our LGN models significantly outperform any version of VGG, or our CNN, at predicting this data. (B.) Log-Likelihoods of observed data given each model as an observer (See Appendix B for details). This analysis allows us to ask how well model-derived distances match human perception over longer distances. These results also show that our LGN models (specifically OnOff) outperform our other model classes.

psychophysical results, butressing the results above.

2.4 Analysis and Extensisions to Realistic Models of Neural Noise

We have presented a new methodology for synthesizing most and least-noticeable distortions from perceptual models, applied this methodology to a set of different models, and tested the resulting predictions by measuring their detectability by human subjects. We show that this methodology provides a powerful form of "Turing test": perceptual measurements on this limited set of model-optimized examples reveal failures that are not be apparent in measurements on a large set of hand-curated examples.

We are not the first to introduce a method of this kind. Wang & Simoncelli, (2008) introduced Maximum Differentiation (MAD) competition, which creates images optimized for one metric while holding constant a competing metric's rating. Our method relies on a Fisher approximation to generate extremal perturbations, and uses the ratio of their empirically measured discrimination thresholds as an absolute measure of alignment to human sensitivity (as opposed to relative pairwise comparisons of model performance). Our method can easily be generalized to incorporate more physiologically realistic noise assumptions, such as Poisson noise, and could potentially be extended to include noise at each stage of a hierarchical model.

We've used this method to analyze the ability of VGG16, a deep convolutional neural network trained to recognize objects, to account for human perceptual sensitivity. First, we find that the early layers of the network are moderately successful in this regard. Second, these layers (Front, Layer 3) surpassed the predictive power of a generic shallow CNN explicitly trained to predict human perceptual sensitivity, but underperformed models of the LGN trained on the same objective. And third, perceptual sensitivity predictions

synthesized from a layer of VGG16 decline in accuracy for deeper layers.

We also showed that a highly structured model of the LGN generates predictions that substantially surpass the predictive power of any individual layer of VGG16, as well as a version of VGG16 trained to fit human sensitivity data (VGG-IQA), or a generic 4-layer CNN trained on the same data. These failures of both the shallow and deep neural networks were not seen in traditional cross-validation tests on the human sensitivity data, but were revealed by measuring human sensitivity to model-synthesized eigen-distortions. Finally, we confirmed that known functional properties of the early visual system (On and Off pathways) and ubiquitous neural computations (local gain control, Carandini & Heeger, (2012)) have a direct impact on perceptual sensitivity, a finding that is buttressed by several other published results (Ballé et al., (2017), Laparra et al., (2017, 2010), Lyu & Simoncelli, (2008), and Malo et al., (2006)).

Most importantly, we demonstrate the utility of prior knowledge in constraining the choice of models. Although the biologically structured models used components similar to generic CNNs, they had far fewer layers and their parameterization was highly restricted, thus allowing a far more limited family of transformations. Despite this, they outperformed the generic CNN and VGG models. These structural choices were informed by knowledge of primate visual physiology, and training on human perceptual data was used to determine parameters of the model that are either unknown or underconstrained by current experimental knowledge. Our results imply that this imposed structure serves as a powerful regularizer, enabling these models to generalize much better than generic unstructured networks.

2.4.1 Predictions Under Poisson Noise Assumptions

In the preceding sections, we made simplifying assumptions about the noise characteristics within our neural network models, assuming homoskedastic gaussian noise. This choice made sense for our initial analysis for several reasons; first, the models we tested all utilized Euclidean distance in their internal representation as a measure of distortion distance, thus implicitly assuming a model of homoskedastic noise, and second, making the noise isotropic allowed us to tease apart how well each model stretches and compresses the perceptual space in line with human perceptual sensitivity without contamination from the effects of a noise model that also reshapes the distortion space. As a model of human physiology and perception, however, our simplifying assumptions deviate from noise distributions observed in the primate visual system. We can reformulate the problem to incorporate Poisson noise by re-deriving the closed-form Fisher Information matrix (from section 3.2.1) under Poisson noise assumptions.

We again return to a set of models that can be expressed by a deterministic (and differentiable) mapping from the input pixels to a mean output firing rate response vector, $\vec{\lambda} = f(\vec{x})$, and with covariance matrix $\Sigma = diag(\lambda)$. The vector of spike counts, $\vec{r}$, on any given observation is a sample from N independent poisson distributions with mean firing rates determined by the elements in $\vec{\lambda}$. The log likelihood for a neural population with independent Poisson variability is:

$$p(\vec{r}|\vec{x}) = \prod_{i=1}^{N} \frac{f_i(\vec{x})^{r_i} e^{-f_i(\vec{x})}}{r_i\ !}$$

And the log likelihood is:

$$\log p(\vec{r}|\vec{x}) = \sum_{i=1}^{N} r_i \; log(f_i(\vec{x})) - f_i(\vec{x}) - log(r_i \; !)$$

Taking the derivative with respect to $\vec{x}$:

$$\frac{\partial}{\partial \vec{x}} \log p(\vec{r}|\vec{x}) = \sum_{i=1}^{N} \frac{(r_i f_i'(\vec{x}))}{f_i(\vec{x})} - f_i'(\vec{x})$$

We can rearrange this, and combine terms such that:

$$\frac{\partial}{\partial \vec{x}} \log p(\vec{r}|\vec{x}) = \sum_{i=1}^{N} \frac{(r_i - f_i(\vec{x}))f_i'(\vec{x})}{f_i(\vec{x})}$$

We can rewrite this in vector notation, substituting the precision matrix Σ^{-1} for $\frac{1}{f(\vec{x})}$:

$$\frac{\partial}{\partial \vec{x}} \log p(\vec{r}|\vec{x}) = \frac{\partial f}{\partial \vec{x}}^T \Sigma^{-1}[\vec{r} - f(\vec{x})]$$

Plugging into equation 3.2, we obtain the following:

$$J[\vec{x}] = \mathbb{E}_{\vec{r}|\vec{x}} \left[\frac{\partial f}{\partial \vec{x}}^T \Sigma^{-1}[\vec{r} - f(\vec{x})][\vec{r} - f(\vec{x})]^T \Sigma^{-1} \frac{\partial f}{\partial \vec{x}} \right]$$

The expectation over $\vec{r}$ of $[\vec{r} - f(\vec{x})][\vec{r} - f(\vec{x})]^T = \Sigma$, by definition. Substituting this back into the equation above:

$$J[\vec{x}] = \frac{\partial f}{\partial \vec{x}}^T \Sigma^{-1} \frac{\partial f}{\partial \vec{x}}$$

That is, the Fisher Information under Poisson noise assumptions induces a locally adaptive metric on the space of images weighted by the precision matrix of the representation. Unlike

the homoskedastic Gaussian noise case, output coefficients with large responses will have larger uncertainty than coefficients with a smaller response, and thus distortions that cause changes to coefficients with large responses will be less detectable than equivalent changes to coefficients with small responses.

We can see the effect that this weighted precision matrix has on the eigen-spectrum of a model by comparison to the eigen-spectrum of the same model under homoeskedastic gaussian noise assumptions. To remind us, the Fisher information in the homoskedastic Gaussian case, $J[\vec{x}]_G$ was defined as:

$$J[\vec{x}]_G = \frac{\partial f}{\partial \vec{x}}^T \frac{\partial f}{\partial \vec{x}}$$

$\frac{\partial f}{\partial \vec{x}}$ can be rewritten as its singular-value decomposition:

$$\frac{\partial f}{\partial \vec{x}} = USV^T$$

And the Fisher Information can thus be rewritten as:

$$J[\vec{x}]_G = VS^2V^T$$

That is, the eigenvectors of $J[\vec{x}]_G$ are equivalent to the right-singular vectors of $\frac{\partial f}{\partial \vec{x}}$, and the eigenvalues of $J[\vec{x}]_G$ are the squared singular values of $\frac{\partial f}{\partial \vec{x}}$.

Returning to the Fisher Information under Poisson assumptions, $J[\vec{x}]_P$, we can similarly rewrite $J[\vec{x}]_P$ as:

$$J[\vec{x}]_P = VSU^T\Sigma^{-1}USV^T$$

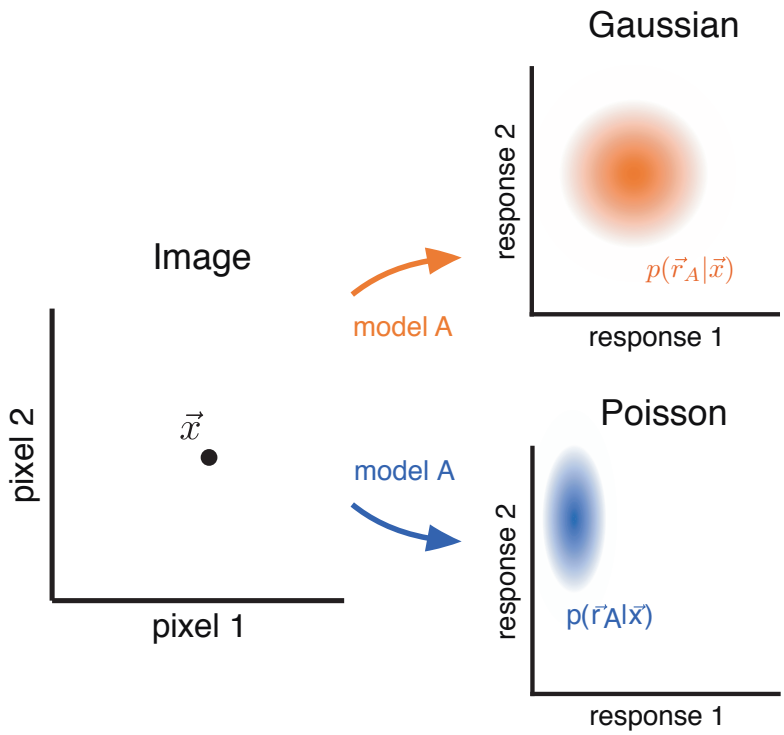

Figure 2.14: Gaussian noise and Poisson noise behave differently, impacting the sensitivity of the model under different noise assumptions. In the gaussian case, the noise variance is indpendent of the mean firing rate, and so the set of detectable distortions around an image always lie on a circular level set (or hyperspherical in higher dimensions) around the mean response to the image. In the Poisson case, the variance of the noise scales with the mean, and so the level set of equivalent sensitivity will be different for different mean responses. Unless the mean values of different neurons are equivalent, the noise cloud in the Poisson case is ellipsoidal (or hyper-ellipsoidal in higher dimensions) and stretched along dimensions with higher response means, decreasing model sensitivity along those directions (by the inverse of the variance of the noise).

If we rename the internal rotation and scaling:

$$A = U^T \Sigma^{-1} U$$

and rewrite $J[\vec{x}]_P$:

$$J[\vec{x}]_P = VSASV^T$$

We see that, unlike in the Gaussian case, the relationship between the singular vectors of $\frac{\partial f}{\partial \vec{x}}$ and $J[\vec{x}]_P$ is much more complicated. In fact, we cannot know a priori if the effects of the matrix A will be to stretch the space in the same directions as the components of $\frac{\partial f}{\partial \vec{x}}$, if it will have counteractive effects to the components of $\frac{\partial f}{\partial \vec{x}}$, or simply no effect at all. We can however, examine the effects empirically.

We utilize the method introduced above for finding the maximum and minimum eigenvector, now under the assumptions of Poisson noise. We find that in general, eigenvectors from models that contain divisive normalization do not change significantly, while eigenvectors from the neural networks lacking divisive normalization change substantially. The simple reason for this is that divisive normalization acts to equalize the variance between neurons, counteracting the effects of the Poisson noise on detectability. We can understand this difference by examining the coefficients of variation for the output response of different models to the same image. Because neurons suppress each other, the relative variance of outputs within normalized models in response to a natural image is much smaller than for non-normalized models, and thus the precision matrix does not reshape the space significantly. For models without local normalization, the relative variance between output coefficients can be much larger, and thus the change from isotropic Gaussian noise to Poisson noise has a large effect on the most and least detectable directions. This helps to

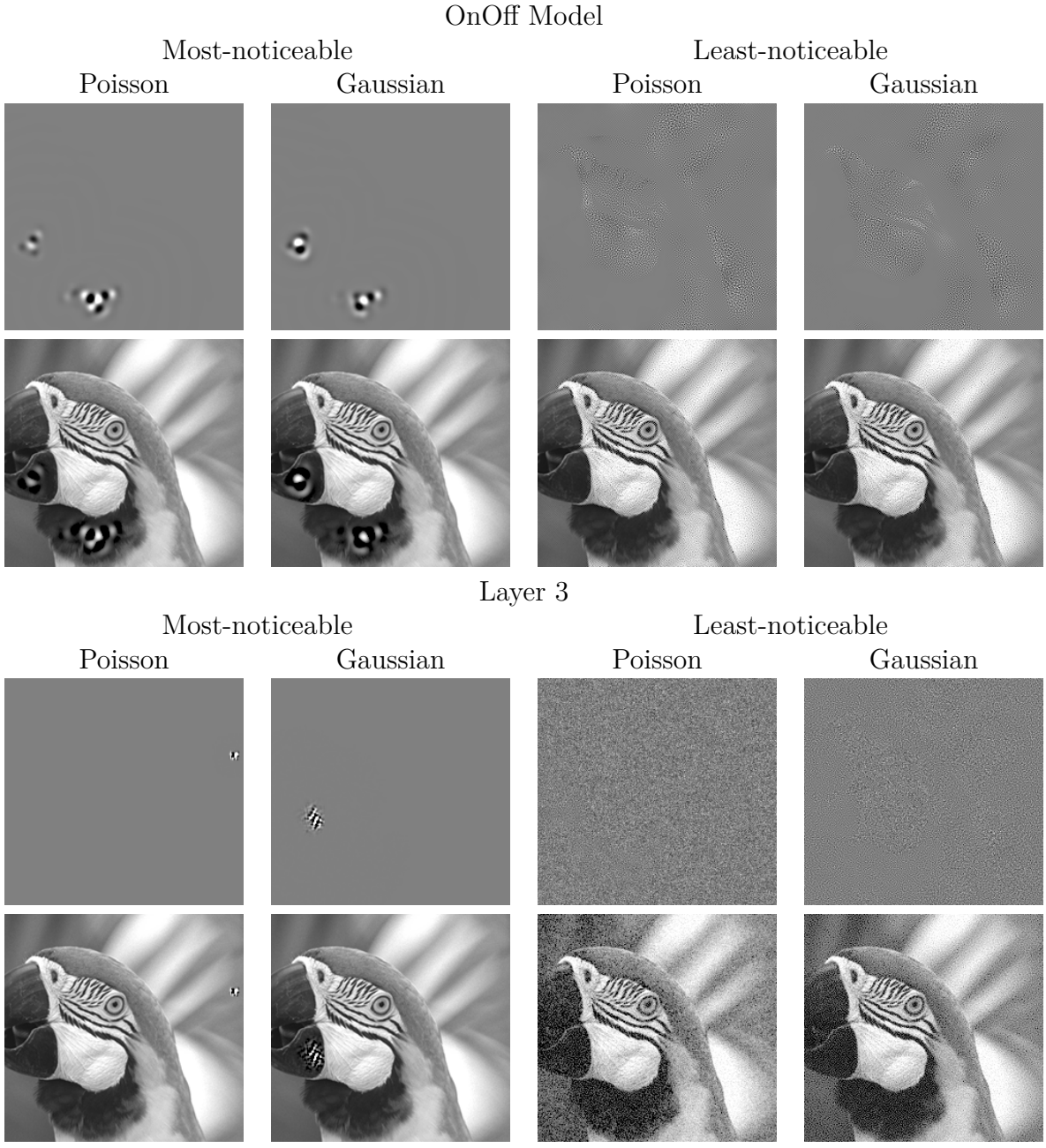

Figure 2.15: Eigen-Distortions for On-Off Model and VGG Layer 3 with Poisson Noise. The predictions from the OnOff model are not substantially modified by changing the noise assumptions. The Predictions from VGG layer 3, however, substantially change when we incorporate Poisson noise assumptions.

explain why, even though we used a simplified noise model, the biological model predictions matched human perception fairly well. A quick glance at the Poisson generated predictions from VGG however shows that under realistic biological noise assumptions, the predictions from VGG get even worse than under our simplified model.

2.4.2 Developing a Poisson Noise Based Distance Metric

If we re-express MSE, or rather the identity transform, $I\vec{x}$, in our Fisher information framework, we can find a simple generalization to computing distance within a space warped by Poisson noise. $\frac{\partial f}{\partial \vec{x}}$ in this case is of course also the identity matrix, I, and thus Fisher Information for the case of MSE reduces to the identity matrix, as discriminability is equivalent in every direction. The eigenvector problem in this case is degenerate.

$$MSE = ||\vec{x} - (\vec{x} + \alpha\vec{u})||_2$$

We can rewrite this in terms of Fisher information:

$$MSE = [\vec{x} - (\vec{x} + \alpha\vec{u})]^T J[f(\vec{x})][\vec{x} - (\vec{x} + \alpha\vec{u})]$$

Or equivalently for the Gaussian case where $J[f(\vec{x})] = I$:

$$MSE = [\alpha\vec{u}]^T[\alpha\vec{u}]$$

For a Poisson noise metric, however, the case is slightly more interesting. The Fisher information matrix under Poisson noise assumptions is equivalent in this case to the precision matrix, Σ^{-1}, a diagonalized version of the inverse of the image pixel values. The equivalent

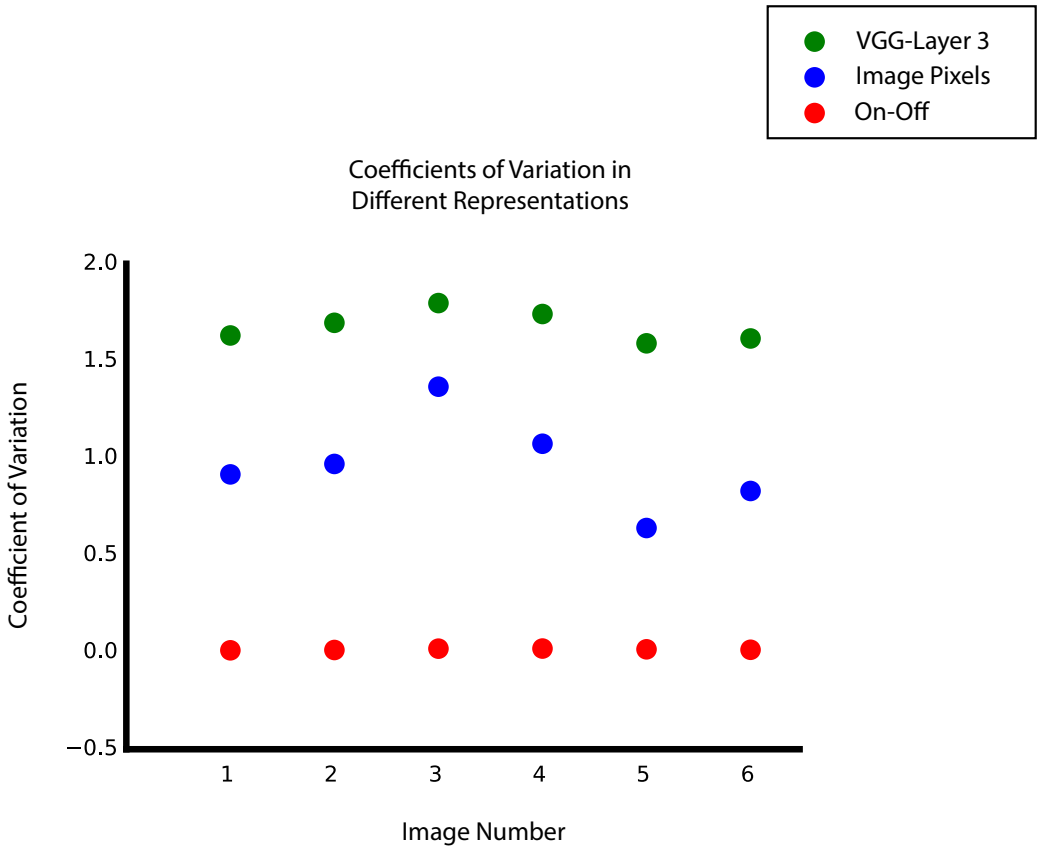

Figure 2.16: Coefficients of Variation within Different Representations. The coefficient of Variation within the OnOff representation is significantly smaller than that of the VGG layer 3 representation. This means that, for the OnOff model, the precision matrix in the Poisson derivation of Fisher Information is close to a scaled version of the identity matrix. This partially explains the relative insensitivity of the OnOff metric to the change from Gaussian to Poisson noise assumptions.

metric in this case is:

$$MSEP = [\vec{x} - (\vec{x} + \alpha\vec{u})]^T \Sigma_{\vec{x}}^{-1} [\vec{x} - (\vec{x} + \alpha\vec{u})]$$

Or, equivalently:

$$MSEP = [\alpha\vec{u}]^T \Sigma_{\vec{x}}^{-1} [\alpha\vec{u}]$$

Where $\Sigma_{\vec{x}}^{-1}$ is the Fisher Information matrix of the undistorted image, $\vec{x}$. This metric can be used as a substitute for measuring distance within a model that assumes Poisson noise.

The eigenvector problem in this case is not degenerate, so we may sample the most- and least-noticeable predictions given this model (see Figure 2.17). This demonstration shows that the Poisson metric, when applied to image pixels, is less sensitive to distortions in high luminance areas of the image. This is predictable from the formulation above. However, at the same time, it is overly sensitive to small distortions low-luminance areas within the image that are hard for humans to see. It is easy to see from this demonstration how the combination of a model like OnOff with Poisson noise would make strong predictions about human sensitivity in both directions.

2.4.3 Models with Equivalent Fisher Information Under Different Noise Assumptions

We may want to evaluate models that incorporate Poisson noise assumptions using the standard psychophysical toolset from signal detection theory, which is constructed under Gaussian noise assumptions. Because Fisher Information depends on both the underlying deterministic transform, and the type of noise within the system, we can trade the effects of one type of noise on model sensitivity with additional deterministic transformations that

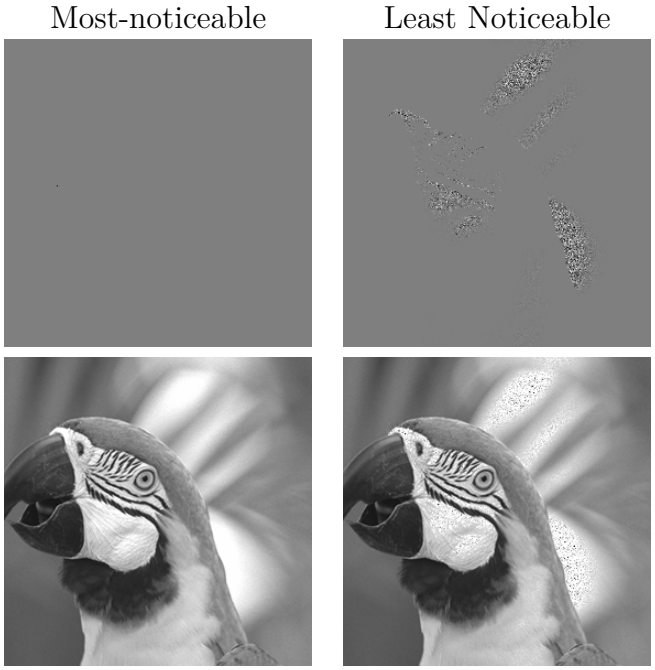

Figure 2.17: Eigen-distortions derived from a Poisson MSE metric show that this simple modification to MSE predicts insensitive directions, but makes worse predictions of sensitive directions than MSE or even our basic LGN models.

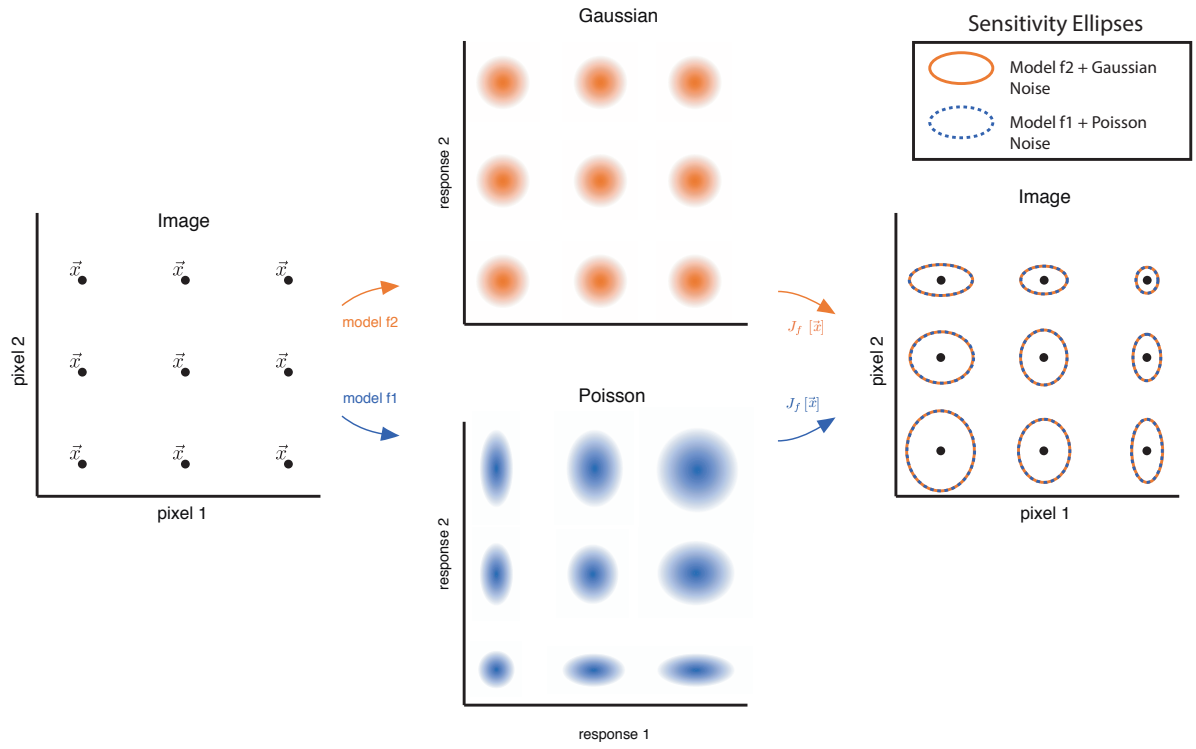

Figure 2.18: Finding a model f_2 under Gaussian noise assumptions with equivalent Fisher Information to model f_1 under Poisson noise assumptions. (See text for details)

have the same effects. We take advantage of this by finding a transformation, f_2 that produces the same Fisher Information under Gaussian noise assumptions as our original transformation, f_1, under Poisson noise assumptions (see Figure 2.18).

$$J[f_2(\vec{x})]_G = J[f_1(\vec{x})]_P$$

For the case where f_1 is the identity matrix times an image, $I\vec{x}$:

$$\frac{\partial f_2}{\partial \vec{x}}^T \frac{\partial f_2}{\partial \vec{x}} = \Sigma^{-1}$$

Where the diagonal elements of Σ^{-1} are the inverse of the pixel values, $\frac{1}{x}$. In the one dimensional case, this is:

$$f_2'(x_i)^2 = \frac{1}{x_i}$$

Or, equivalently:

$$f_2'(x_i) = \frac{1}{\sqrt{x_i}}$$

Taking the integral (and disposing of the constant):

$$f_2(x_i) = 2\sqrt{x_i}$$

That is, replacing $f_1 = I$ with $f_2 = 2\vec{x}^{\frac{1}{2}}$, results in an equivalent Fisher information matrix under Gaussian noise assumptions for f_2 and Poisson noise assumptions for f_1.

$$\frac{\partial f_2}{\partial \vec{x}}^T \frac{\partial f_2}{\partial \vec{x}} = (\Sigma^{-\frac{1}{2}})^T(\Sigma^{-\frac{1}{2}}) = \Sigma^{-1}$$

Using this, we can utilize distances in the space of f_2 to fit psychophysical data using the same Gaussian assumptions commonly used, while simultaneously matching distances, in the non-Euclidean space of f_1.

The inverse of this, in which we wish to find a function f_1 with Poisson noise that has equivalent Fisher Information to a function f_2 with gaussian noise (see Figure 2.19), is not uniquely constrained, but a family of solutions can be easily found (see Figure 2.19).

Again, we have:

$$J[f_2(\vec{x})]_G = J[f_1(\vec{x})]_P$$

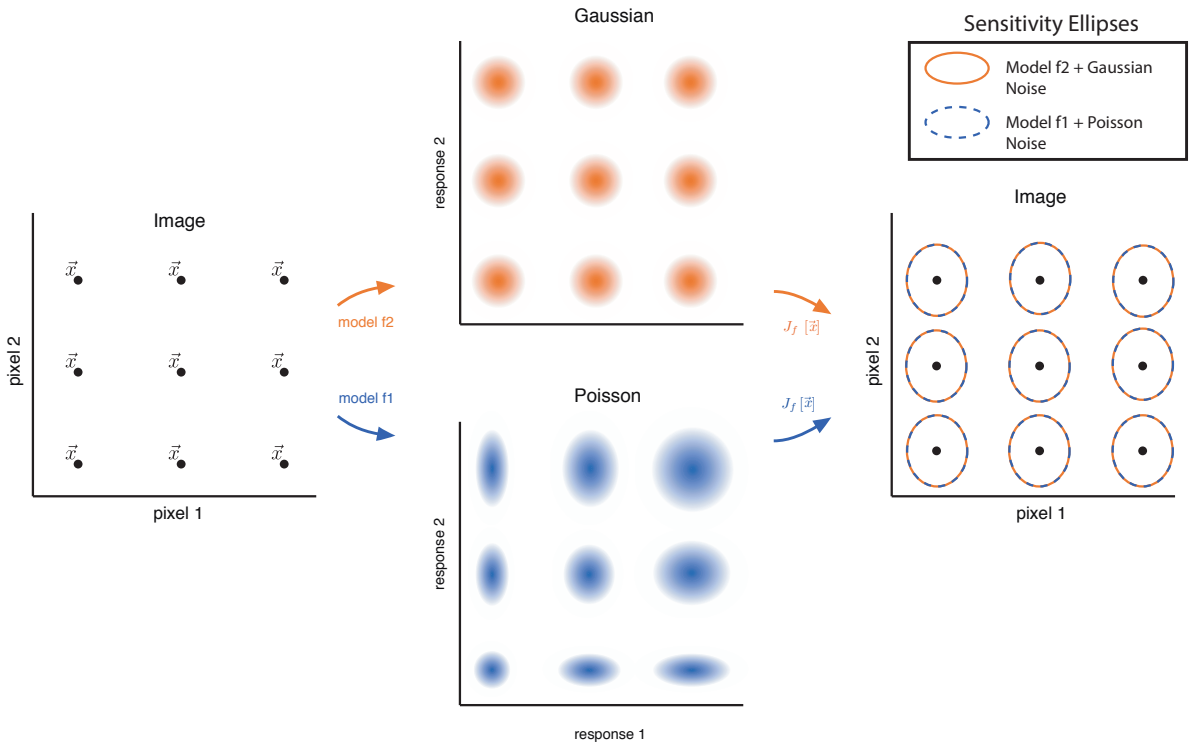

Figure 2.19: Finding a model f_1 under Poisson noise assumptions with equivalent Fisher Information to model f_2 under Gaussian noise assumptions. (See text for details)

However, in this case, let us assume that f_2 is the identity matrix times an image, $I\vec{x}$:

$$I = \frac{\partial f_1}{\partial \vec{x}}^T \Sigma^{-1} \frac{\partial f_1}{\partial \vec{x}}$$

In the one dimensional case: this is:

$$\frac{(f_1'(x))^2}{f_1(x)} = 1$$

This is a first-order nonlinear ordinary differential equation that is under constrained and thus, there are an infinite family of functions that satisfy this condition. We will solve for one condition assuming all constants $c = 0$. Since we have all terms that depend on $f(x)$ on one side, we can take the integral of both sides:

$$\int \frac{f_1'(x)dx}{\sqrt{f_1(x)}} = \int 1dx$$

Which is equal to:

$$2\sqrt{f_1(x)} = x$$

Rearranging:

$$f_1(x) = \frac{1}{4}x^2$$

That is, replacing $f_2 = I$ with $f_1 = \frac{1}{4}\vec{x}.^2$, where $.^2$ indicates an element wise squaring operation, results in an equivalent Fisher information matrix under Poisson noise assumptions for f_1 and Gaussian noise assumptions for f_2.

$$1 = \frac{(\frac{1}{2}x)^2}{\frac{1}{4}x^2} = \frac{(f_1'(x))^2}{f_1(x)}$$

The transferability of Fisher Information between two different deterministic models under different stochastic assumptions highlights a key benefit to our method of constructing neural perceptual quality metrics as a combination of deterministic transformations plus additive noise. The contributions to the sensitivity of these metrics that are properties of the noise can alternatively be traded off for specific transformations preceding the noise, allowing for model simplification and extension.

Generating Eigen-distortions for single-scale SSIM

Because SSIM is still the most widely used perceptual similarity metric, we wanted to compare the predictions of our models to its predictions. However, SSIM is composed from the product of three correlations, and cannot be decomposed into a form that is compatible with our eigen-distortion analysis, as it can not be reduced to a simple transformation with additive noise. While this is unsatisfactory from our perspective, as we would like to compare SSIM's predictions to the rest of our models, it highlights one of the main advantages of our models, and one of the main disadvantages of SSIM. Our metrics are optimized to model perceptual distances under Euclidean assumptions, allowing us to separate the action of the function and the assumed noise model. In addition, we can easily extend our metrics to include more processing stages and more complicated forms of noise, and we can test those extended models with closed form solutions for their Fisher information. Our models also allow us to ask to what degree different elements of the functional architecture of the visual system contribute to capturing human perceptual sensitivity, allowing us to tie together physiology and perception. It is not clear how to extend SSIM beyond its current iteration.

Chapter 3

Perceptually Optimized Image Rendering

3.1 The Problem of Optimal Image Rendering

A general goal in designing a pipeline for the capture and display of photographic images is to remain as faithful to the original source as possible, minimizing distortions introduced by the sensor, coding, transmission, or display processes. If images are meant for presentation to human observers, distortion should be measured accordingly, penalizing errors that are most visually noticeable and/or disturbing, while permitting those that are perceptually unnoticeable. This strategy is most evident in the handling of color, in which both sensors and displays are designed so as to accurately capture and render the three-dimensional subspace of wavelengths relevant for human trichromatic visual representation, while allowing significant distortion outside of this subspace.

Arguably the most significant limitations of current sensors and displays are with regard to dynamic range. Early digital sensors were restricted to capturing a limited luminance range, and were unable to adequately capture the majority of realistic natural scenes, which

contain luminances spanning up to roughly 20 orders of magnitude. In contrast, the human visual system is capable of sensing fixed scenes with a range of over 5 orders of magnitude in real time, up to 8 orders of magnitude in the photopic regime when the effects of extended temporal adaptation mechanisms are incorporated Hoefflinger, (2007), and up to 14 orders of magnitude when including the scotopic and mesopic regimes (see Fig. 3.1). The dynamic range of sensors has steadily improved, and current sensors (often augmented with software solutions that fuse images captured at different exposures) are capable of acquiring images with luminance ranges approximating those of human vision. Despite this, even the best display devices are limited to a significantly lower dynamic range than these sensors can capture.

The simplest solution to the problem of displaying high dynamic range (HDR) images on a low dynamic range (LDR) rendering device is to linearly rescale the luminance values recorded by the sensor into the display's reproducible range of luminances. This, however, produces images that look nothing like the original scene – typically all of the low-luminance information is lost. A variety of tone-mapping methods have been proposed to solve this problem by nonlinearly remapping the intensities of the original image into the output range, in a way that least interferes with the visual appearance of the original scene. Most of these are based on heuristics, and require manual parameter adjustment for best results. In addition, many displays introduce constraints other than global luminance range, such as restriction to discrete luminance levels (i.e. halftoning), maximal average power consumption, and interactions between pixel values over space or time. Separate methods have been developed for solving each of these problems.

Perceptual optimization of tone mapping was introduced in a seminal paper by Tumblin and Rushmeier, who proposed the selection of a tone mapping transformation from HDR

images to LDR displays so as to best match the appearance of the original scene Tumblin & Rushmeier, (1993). A variety of tone mapping papers have followed this framework (see for instance Ferwerda et al., (1996), Mantiuk & Kerofsky, (2008), Pattanaik et al., (1998, 2000), and Tumblin et al., (1999)). These methods are dependent on the parametric function used as a tone mapping operator, which restricts the space of possible solutions: A given functional form may not be able to achieve a perceptually optimal solution, or may only work satisfactorily for a particular type of rendering constraint.

Here, we formulate a more general solution for perceptually accurate rendering, directly optimizing the rendered image so as to minimize perceptual differences with the light intensities of the original scene, subject to all constraints imposed by the display (Fig. 3.1). This constrained optimization formulation relies on four ingredients: knowledge of the original scene luminances (or calibration information that allows calculation of those luminances), a measure of perceptual similarity between images, knowledge of the display constraints, and a method for optimizing the image to be rendered. We use a model of perceptual similarity, loosely based on the transformations of the early stages of the Human Visual System (specifically, the retina and LGN), that has previously been fit to a database of human psychophysical judgments. Because this model is continuous and differentiable, our method can be efficiently solved by first-order constrained optimization techniques. We show that the solution is well-defined and general, and therefore represents a framework for solving a wide class of rendering problems.

In section 3.3.1, we optimize images captured under differing acquisition conditions for rendering on the same display. We show one result per experiment – more images can be found at `http://www.cns.nyu.edu/~lcv/perceptualRendering/`. We start with calibrated images, where the original scene luminances are known. We also deal with the

more common scenario in which the exact luminances of the original scene are unknown (the tone mapping problem). In this scenario, we have to make some educated guesses about the luminance range of the original scene, and we demonstrate the effect that different assumptions have on the optimized images. Moreover, we take advantage of these effects to solve other image processing problems, such as detail enhancement and haze removal, by manipulating these source assumptions. For each of these tasks, we compare the results with state-of-the-art algorithms designed to solve each specific case. In section 3.3.4, we optimize images to be displayed under differing display restrictions, including luminance limited displays, power limited displays, and displays restricted to a small set of output values. Finally, we analyze the effect that each component of our perceptual measure has on the quality of our optimized images.

3.2 Framework Development

3.2.1 Optimal Rendering Framework

Optimally rendering an image, $\boldsymbol{I}$, on a given display means displaying it in such a way that it remains faithful to the human perception of the original scene, $\boldsymbol{S}$. Here, $\boldsymbol{S}$ and $\boldsymbol{I}$ are vectors representing the luminances of all pixels in the respective images. We formalize this as a constrained optimization problem:

$$\hat{I}_{\mathcal{C}}(\boldsymbol{S}) = \arg\min_{\boldsymbol{I}} D(\boldsymbol{S}, \boldsymbol{I}), \quad \text{s.t. } \boldsymbol{I} \in \mathcal{C}, \tag{3.1}$$

where $D(\cdot, \cdot)$ is a measure of human perceptual dissimilarity, and $\mathcal{C}$ is the set of all images that can be rendered on the display. This formulation can express many well-known rendering problems, such as tone mapping or dithering, which differ only in the specification of $\mathcal{C}$.

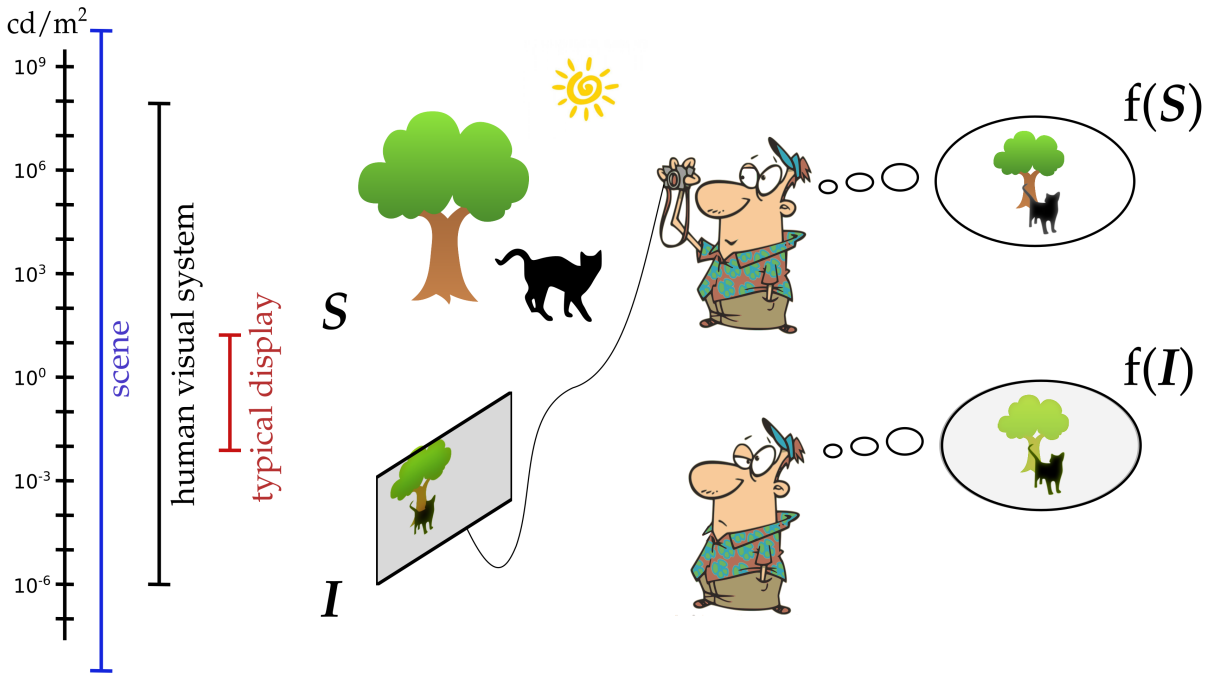

Figure 3.1: Perceptually optimized rendering framework. When we view a real-world scene, the luminances, specified by a vector $\boldsymbol{S}$, give rise to an internal perceptual representation $f(\boldsymbol{S})$. While luminances in the real world can range from complete darkness ($0\,cd/m^2$) to extremely bright (e.g., midday sun, roughly $10^9\,cd/m^2$), a typical display can generate a relatively narrow range of roughly 5 to $300\,cd/m^2$. The optimization goal is to adjust luminances $\boldsymbol{I}$ generated by the display, so as to minimize the difference between the perceptual representations, $f(\boldsymbol{S})$ and $f(\boldsymbol{I})$, while remaining within the set of images that can be generated by the display.

In general, the optimization problem expressed in Eq. (3.1) cannot be solved analytically, and thus we will not obtain an explicit function to compute $\hat{I}_{\mathcal{C}}(\boldsymbol{S})$, given $\boldsymbol{S}$ and $\mathcal{C}$. Instead, we choose a perceptual measure that is differentiable with respect to $\boldsymbol{I}$, and use modern high-dimensional optimization tools to numerically solve for $\hat{I}_{\mathcal{C}}(\boldsymbol{S})$. Specifically, we descend the objective function, alternating between minimizing the perceptual distance, and projecting the image back onto the constraint set. Specific formulations for different example problems can be found online at `http://www.cns.nyu.edu/~lcv/perceptualRendering/`.

We follow a principled, two-step approach to quantify perceptual distance. Rather than defining a perceptual distance directly (as in SSIM (Wang et al., 2004), for example), we first define a nonlinear *perceptual transform* $f(\cdot)$, which approximates the computations performed within the early stages of the human visual system. We apply this to both the original scene luminances, $\boldsymbol{S}$, and the rendered image, $\boldsymbol{I}$, and then measure the distance between $f(\boldsymbol{S})$ and $f(\boldsymbol{I})$. We refer to this casually as a "metric" (as is common in the image quality assessment literature), even though it is not guaranteed to satisfy all requirements of the mathematical definition of a metric. Specifically, it is symmetric and yields a value zero for identical images, but for some parameter values the transformation can discard information (allowing it to produce a zero distance for non-identical images), and it also may not satisfy the triangle inequality.

3.2.2 Development of a Multiscale Metric

Normalized Laplacian pyramid model

Figure 3.2 illustrates the components of the perceptual transform, which we call the Normalized Laplacian Pyramid (NLP), a multi-scale nonlinear representation. We developed this particular multi-scale metric in order to account for real-world viewing conditions. In

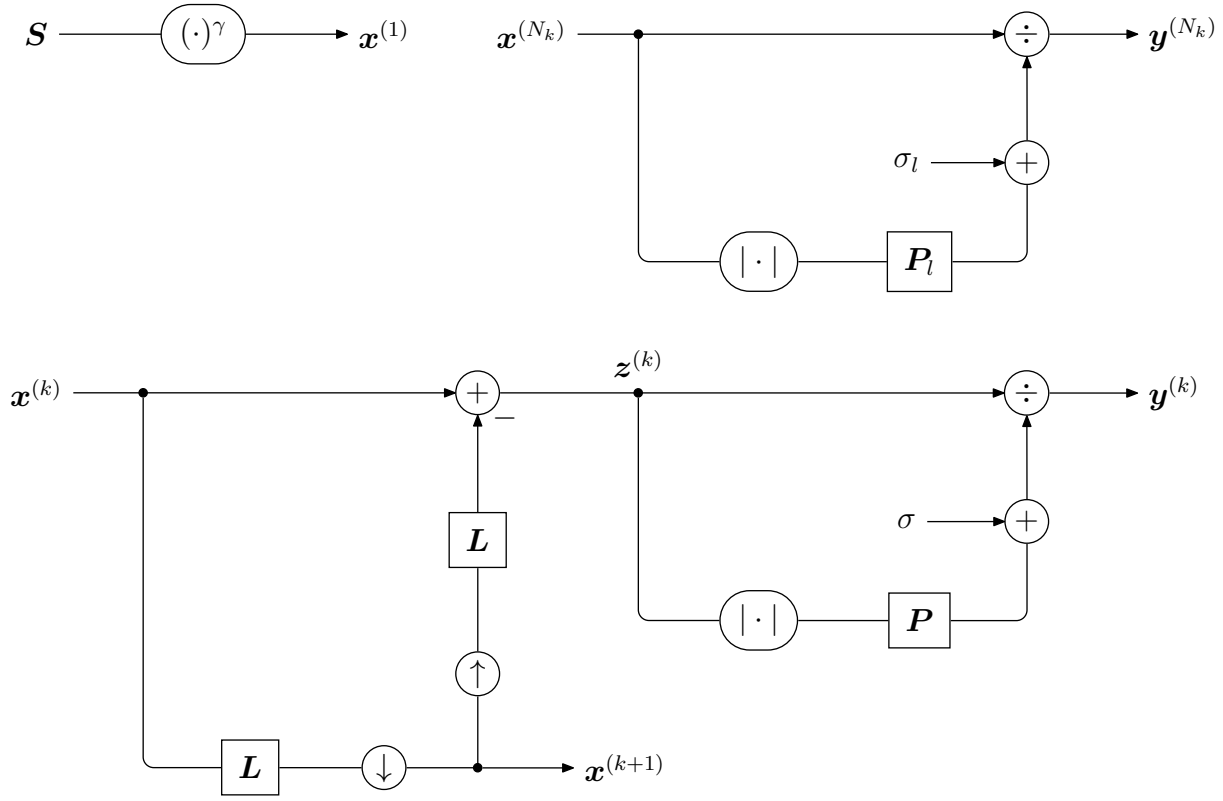

Figure 3.2: Perceptual transform $f(S)$, constructed as a Normalized Laplacian Pyramid (NLP) (Laparra et al., 2016). Scene luminances $\boldsymbol{S}$ (in cd/m^2) are first transformed using a power function (top left). The transformed luminance image is then decomposed into frequency channels, using the recursive implementation of the Laplacian Pyramid (Burt & Adelson, 1983). Each channel $\boldsymbol{z}^{(k)}$ is then divided by a weighted sum of local amplitudes (computed with lowpass filter $\boldsymbol{P}$) plus a constant, σ. The final lowpass channel $\boldsymbol{x}^{(N_k)}$ is also normalized, but with distinct parameters (top right). Symbols $\uparrow$ and $\downarrow$ indicate upsampling and downsampling by a factor of 2, respectively.

the real-world, viewers will not always see images displayed on a screen from a fixed viewing distance, and thus a functionally useful metric needs to work over many viewing distances at once. The single-scale metrics developed and explored in chapters 1 and 2, while not appropriate for this real-world application as constructed, did inspire many of the choices we made in construction of this metric, i.e. it is a metric that mimics the operations of the early stages of the human visual system. This representation is inspired by a model for responses of the lateral geniculate nucleus (LGN) (Mante et al., 2008), and includes contrast gain control mechanisms. This transform bears some resemblance to previously published image metrics that utilize local normalization but differs in motivation, structure, and implementation (Laparra et al., 2016; Mittal et al., 2012, 2013; Teo & Heeger, 1994a; Wang et al., 2004).

Local Mutual information

We view the local luminance subtraction and contrast normalization seen in retinal and LGN computation as a means of reducing redundancy in natural images (as described in chapter 1). Most of the redundant information in natural images is local, and can be captured with a Markov model. That is, the distribution of an image pixel (x_i) conditioned on all others is well approximated by the conditional

$$p(x_i|\boldsymbol{x}_{Ni}), \tag{3.2}$$

where $\boldsymbol{x}_{Ni}$ is the vector of pixels in its immediate neighborhood. This redundancy can be removed by a parametric estimate of a statistic of the central pixel, gathered from its neighbors.

In a initial version of the model, we estimated the normalization parameters from a

large set of undistorted images only (Laparra et al., 2016). This formulation allowed us to build an architecture inspired by the computations of the early visual system and to use a statistical criterion to select the local gain control parameters. Specifically, the weights used in computing the gain signal were chosen so as to minimize the conditional dependency of neighboring transformed coefficients.

In this model, an image is first decomposed by a recursive partition into frequency channels, as in the Laplacian Pyramid (Burt & Adelson, 1983), mimicking the center-surround receptive fields found in retina (and LGN):

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D}\boldsymbol{L}\boldsymbol{x}^{(k)}, \qquad k \in \{1, \ldots, N_k - 1\}, \tag{3.3}$$

$$\boldsymbol{z}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{L}\boldsymbol{U}\boldsymbol{x}^{(k+1)}, \tag{3.4}$$

$$\boldsymbol{z}^{(N_k)} = \boldsymbol{x}^{(N_k)}, \tag{3.5}$$

where $\boldsymbol{D}$ and $\boldsymbol{U}$ indicate down/up-sampling by a factor of two, respectively (figure 3.2). For the filtering operation $\boldsymbol{L}$, we apply a spatially separable 5-tap filter, $(0.05, 0.25, 0.4, 0.25, 0.05)$, as originally specified in (Burt & Adelson, 1983).

Within each channel, each coefficient is divided by a weighted local sum of the element-wise amplitudes (absolute values) plus a constant:

$$y_i = z_i / f_C(\boldsymbol{z}_{Ni}; \sigma, \boldsymbol{p}). \tag{3.6}$$

As an estimate of the local amplitude of a coefficient, f_c, at a given scale, k, we used a linear combination of rectified neighbors:

$$f_C(\boldsymbol{z}_{Ni}^{(k)}; \sigma^{(k)}, \boldsymbol{p}^{(k)}) = \sigma^{(k)} + \sum_{j \in Ni} p_j^{(k)} \left| z_j^{(k)} \right|, \tag{3.7}$$

where $\boldsymbol{p}^{(k)}$ is the vector of non-negative weights, and $\sigma^{(k)}$ is a positive-valued constant, such that f_C is guaranteed to be positive for all neighborhoods, avoiding division by zero. For each scale, the constant is set to the average of the absolute value:

$$\sigma^{(k)} = \frac{1}{N_s^{(k)}} \sum_{i=1}^{N_s^{(k)}} \left| z_i^{(k)} \right|, \tag{3.8}$$

where $N_s^{(k)}$ is the number of coefficients in the subband at scale k. The weight vector is chosen as the solution of the optimization problem:

$$\boldsymbol{p}^{(k)} = \arg\min_{\boldsymbol{p}} \sum_{i=1}^{N_s^{(k)}} \left(\left| z_i^{(k)} \right| - f_C(\boldsymbol{z}_{Ni}^{(k)}; \sigma^{(k)}, \boldsymbol{p}) \right)^2. \tag{3.9}$$

Our final measure of distance is a simple extrapolation of the distance measure we have been utilizing for our single scale metrics. We take the mean squared error across within each scale, and then average across scales.

$$D(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sqrt{N_s^{(k)}}} \left\| \boldsymbol{y}^{(k)} - \tilde{\boldsymbol{y}}^{(k)} \right\|_2 \tag{3.10}$$

where $\boldsymbol{y}^{(k)}$ and $\tilde{\boldsymbol{y}}^{(k)}$ denote vectors containing the transformed reference and distorted image data, respectively. This distance metric implicitly gives more weight to lower frequency coefficients (of which there are fewer, due to subsampling).

We showed that this metric performed at or above the state of the art on several human databases of perceptual quality assessment despite being fit to image statistics and not to human perceptual data (See Appendix C) (Laparra et al., 2016).

Figure 3.4 illustrates the reduction of redundant information at each stage of the model. Each image shows the empirical pairwise mutual information between a given coefficient

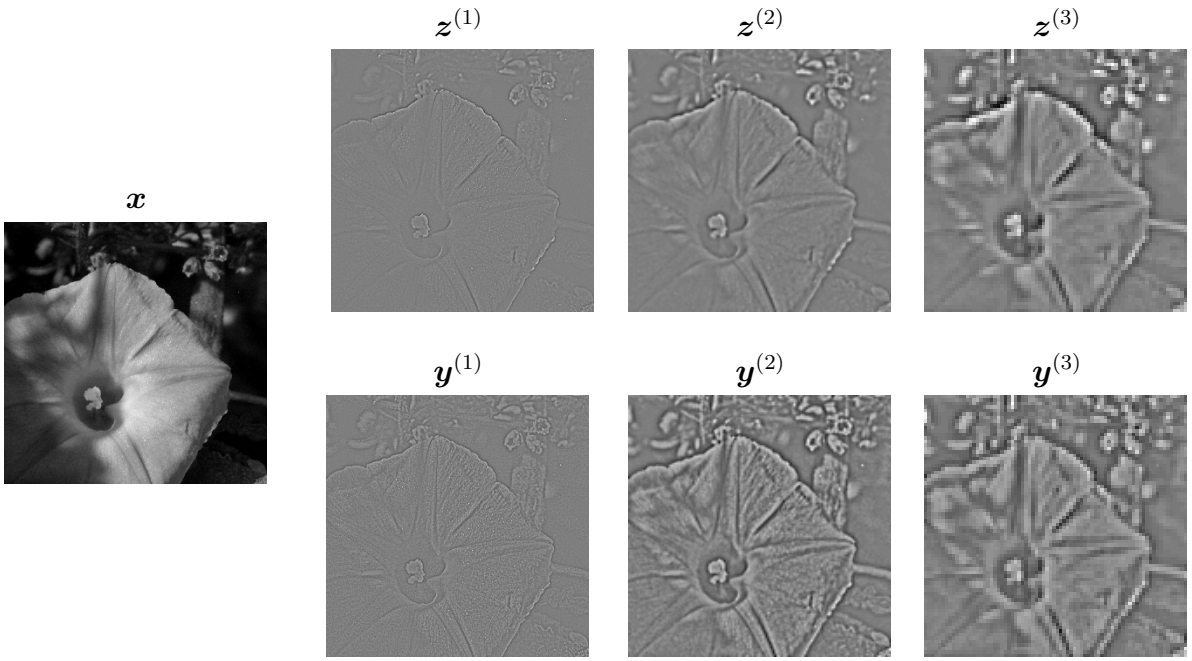

Figure 3.3: Representation of an example image. $\boldsymbol{x}$ is the original image (left). $\boldsymbol{z}$ is the decomposition of the image using the Laplacian pyramid (three scales shown), each image corresponding to a different scale. Note that the Laplacian pyramid includes downsampling in each scale. The examples shown here have been upsampled for visualization purposes. $\boldsymbol{y}$ are the corresponding locally contrast-normalized images.

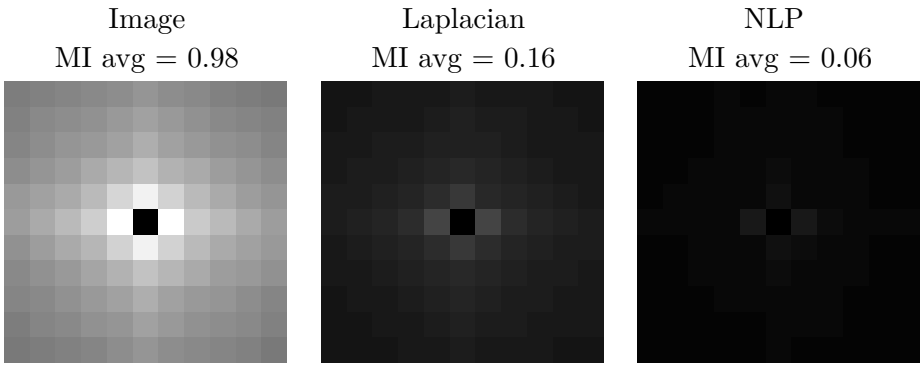

Figure 3.4: Local mutual information between coefficients and their spatial neighbors within an 11×11 local region, for three representations (image pixels, Laplacian pyramid subband, normalized Laplacian pyramid subband). Brightness is proportional to the mutual information between a central coefficient and the neighbor at that relative location. Values are estimated from one million image patches. The average mutual information over the whole neighborhood is given above each panel.

(central pixel of each image) and each of its neighbors. Mutual information has been computed using one million samples from the reference images in the TID database (Ponomarenko et al., 2009). The figure reports the results for the first scale – results for the other scales are similar. The information reduction from both stages of processing is seen to be quite substantial – a factor of roughly six and three, respectively.

Improving this metric by training on a psychophysical objective

In order to improve the performance and generality of this model, first published in (Laparra et al., 2016), we attempted to leverage the power of our previous approaches to the problem, by fitting a model with structure inspired by both the statistics of natural images, as well as the early visual system, to maximize a psychophysical objective. All parameters of the perceptual transform and metric were optimized to best explain human perceptual ratings of distorted images in a public database of grayscale images (Ponomarenko et al., 2009). Specifically, we chose parameters to maximize the correlation between the mean opinion

scores from the human observers and the distance computed by the metric (as in chapter 2). Modifications to the model are shown below.

Here, we adapt this model to operate directly on luminances (in cd/m^2, rather than values that have been gamma-adjusted for a particular display), which provides a standardized set of units for defining constraints on acquisition and display.

Luminances are first transformed element-wise using a power law, which approximates the transformation of light to response of retinal photoreceptors:

$$\boldsymbol{x}^{(1)} = \boldsymbol{S}^{\gamma}. \tag{3.11}$$

The optimized exponent for the front-end nonlinearity was $\gamma = \frac{1}{2.6}$. This initial nonlinear transformation is followed by a recursive partition into frequency channels, as in the Laplacian Pyramid (Burt & Adelson, 1983), just as in the original model:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D}\boldsymbol{L}\boldsymbol{x}^{(k)}, \qquad k \in \{1, \ldots, N_k - 1\}, \tag{3.12}$$

$$\boldsymbol{z}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{L}\boldsymbol{U}\boldsymbol{x}^{(k+1)}, \tag{3.13}$$

$$\boldsymbol{z}^{(N_k)} = \boldsymbol{x}^{(N_k)}, \tag{3.14}$$

Within each channel, each coefficient is divided by a weighted local sum of the element-wise amplitudes (absolute values) plus a constant:

$$\boldsymbol{y}^{(k)} = \boldsymbol{z}^{(k)} \oslash \left(\sigma + \boldsymbol{P}|\boldsymbol{z}^{(k)}|\right), \tag{3.15}$$

where $\boldsymbol{P}$ indicates convolution with a filter, and $\oslash$ indicates point-wise division. All bandpass channels and the highpass channel share the same parameters $\boldsymbol{P}$ and σ, whereas the

lowpass ($k = N_k$) has its own parameter set, $\boldsymbol{P}_l$ and σ_l. This function is a simplified variant of *divisive normalization*, used to describe the responses of neurons in different parts of the visual system (Carandini & Heeger, 2012; Heeger, 1992; Schwartz & Simoncelli, 2001).

For bandpass channels, the additive constant was $\sigma = 0.17$, and the local weighting functions $\boldsymbol{P}$ were filters with 5×5 support, with values:

$$\boldsymbol{P} = \begin{bmatrix} 4 & 4 & 5 & 4 & 4 \\ 4 & 3 & 4 & 3 & 4 \\ 5 & 4 & 5 & 4 & 5 \\ 4 & 3 & 4 & 3 & 4 \\ 4 & 4 & 5 & 4 & 4 \end{bmatrix} \cdot 10^{-2}. \tag{3.16}$$

The parameters for the lowpass channel were $\boldsymbol{P}_l = \mathbf{1}$ (the identity) and $\sigma_l = 4.86$. Optimized exponents for the metric were $\alpha = 2.0$ and $\beta = 0.6$. The NLP coefficients of all channels $\boldsymbol{y}^{(k)}$ combined represent the response of the perceptual transform:

$$f(\boldsymbol{S}) = \left\{ \boldsymbol{y}^{(k)}; k = 1, \ldots, N_k \right\}. \tag{3.17}$$

We wished to see if we could improve upon MSE measured across all channels as the final distance metric in this space. Figure 3.5 illustrates the construction of the metric that we employed to do so. We first computed the L_α-norm of the differences between NLP coefficients within each frequency channel (that is, we raise the absolute value of each coefficient difference to the power α, sum over the entire channel, and take the αth root). These values are then combined across channels using an L_β-norm, to yield the final NLP distance (NLPD):

$$D(\boldsymbol{S}, \boldsymbol{I}) = \left[\frac{1}{N_k} \sum_{k=1}^{N_k} \left(\frac{1}{N_c^{(k)}} \sum_{i=1}^{N_c^{(k)}} \left| y_i^{(k)} - \tilde{y}_i^{(k)} \right|^\alpha \right)^{\frac{\beta}{\alpha}} \right]^{\frac{1}{\beta}}, \tag{3.18}$$

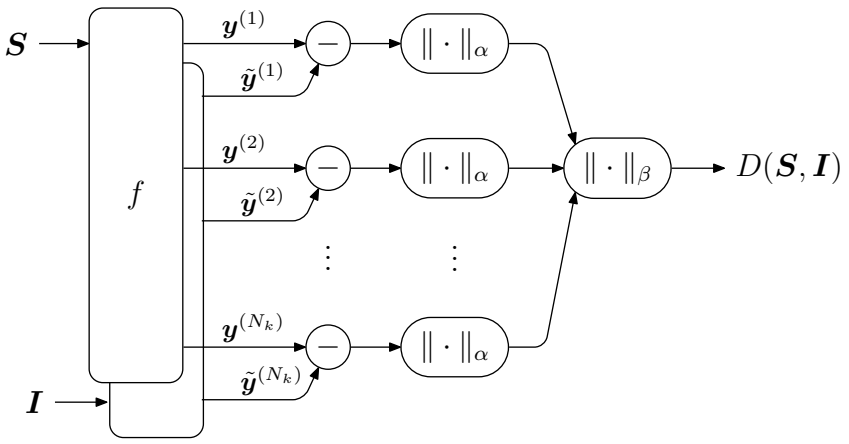

Figure 3.5: Construction of the Normalized Laplacian Pyramid Distance (NLPD) measure. Two images are transformed by $f(\cdot)$ to a perceptual representation, yielding two NLPs (see Fig. 3.2). We compute the α-norm over the vector of differences for each frequency channel, and then combine these over channels using a β-norm. For all rendering results, we use $\alpha = 2.0$ and $\beta = 0.6$, which are optimized to fit the human perceptual ratings of distorted images reported in (Ponomarenko et al., 2009).

where $\tilde{y}_i^{(k)}$ indicates the kth subband arising from the displayed image $\boldsymbol{I}$ (i.e. $f(\boldsymbol{I}) = \{\tilde{\boldsymbol{y}}^{(k)}; k = 1, \ldots, N_k\}$) and $N_c^{(k)}$ is the number of coefficients in that subband. A similar summation model has been employed in previous perceptual quality metrics (Laparra et al., 2010; Watson, 1993). We optimized the parameters, α and β, alongside the rest of our parameters. This optimized distance metric allows us to account for differences in weighting within and across scales.

In Appendix C, we show that the performance of this extended and optimized version of the NLP metric surpasses that of state-of-the-art image quality metrics, as well as our original version. This metric, with parameters held fixed at their optimized values, was used to optimize all of the rendering results presented below.

3.3 Application of the Rendering Framework

3.3.1 Varying Image Acquisition Conditions

We performed a set of experiments to test the capabilities of our optimization framework over different image acquisition conditions. We begin with calibrated images, for which we know the the exact luminance values (in cd/m^2) of the original scene. We then move on to uncalibrated images, for which we need to make an assumption about the luminance values in the original scene. Finally, we close this section by demonstrating that the method is stable and flexible enough that it can be used to solve other rendering problems, such as haze removal and artificial detail enhancement.

Each example requires us to minimize the perceptual distance with respect to the rendered image $\boldsymbol{I}$, subject to the display constraints. In general, this is accomplished by alternating between projection onto the constraint set and minimization of the distance using the Adaptive Moment Estimation (Adam) algorithm (Kingma & Ba, 2014). The gradient of the perceptual distance with respect to $\boldsymbol{I}$ is described in appendix **??**. Implementation of the derivatives, along with additional optimized examples, are provided on the project webpage `http://www.cns.nyu.edu/~lcv/perceptualRendering/`. All images presented here are intended for viewing on a display with luminance ranging from 5 to $300\,cd/m^2$, and a gamma value of 2.2. Computation time scales linearly with the size of the image. When optimized on a Tesla K40 GPU card, it takes approximately 1 second per 10000 pixels (i.e. an image of 1000×1000 requires less than 2 minutes).

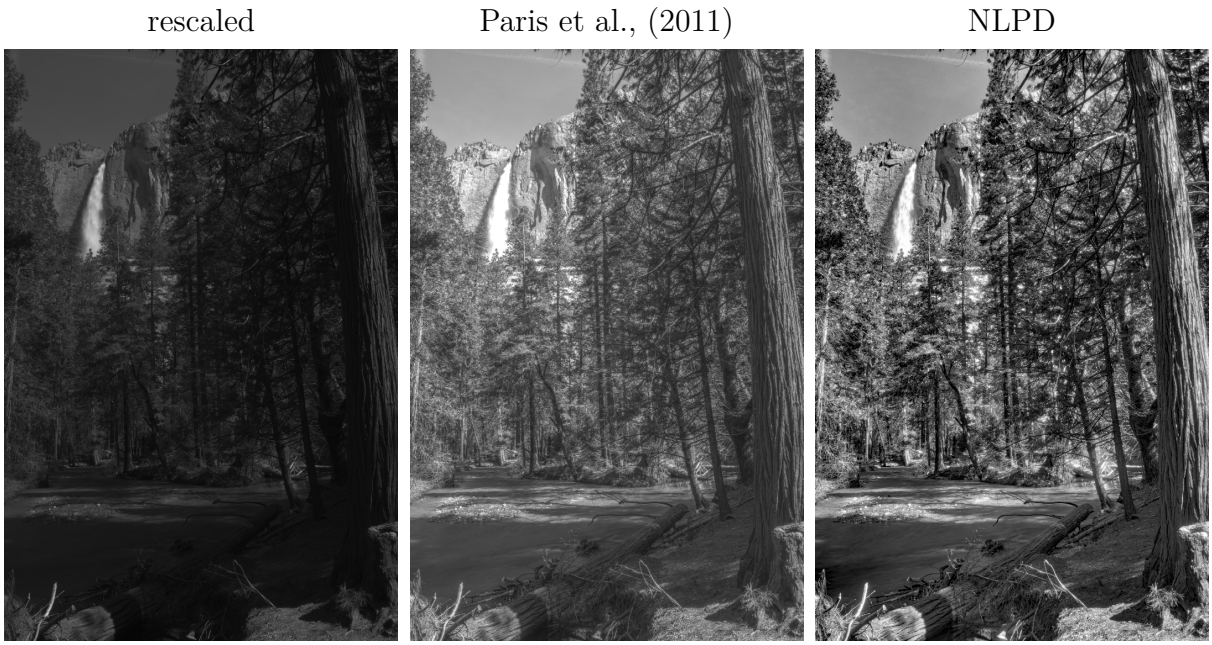

Figure 3.6: Rendering of a calibrated HDR image on a display with a limited luminance range. The scene luminances for this image spanned the range from $S_{\min} = 0.78\,cd/m^2$ to $S_{\max} = 16200\,cd/m^2$, whereas the display luminances are assumed to lie between $5\,cd/m^2$ and $300\,cd/m^2$. Left: the image rendered by linear rescaling of luminance values into the display range. Center: the image rendered using a state-of-the-art tone mapping algorithm (Paris et al., 2011). Right: the image rendered using the proposed method of minimizing the NLPD metric subject to the display constraints.

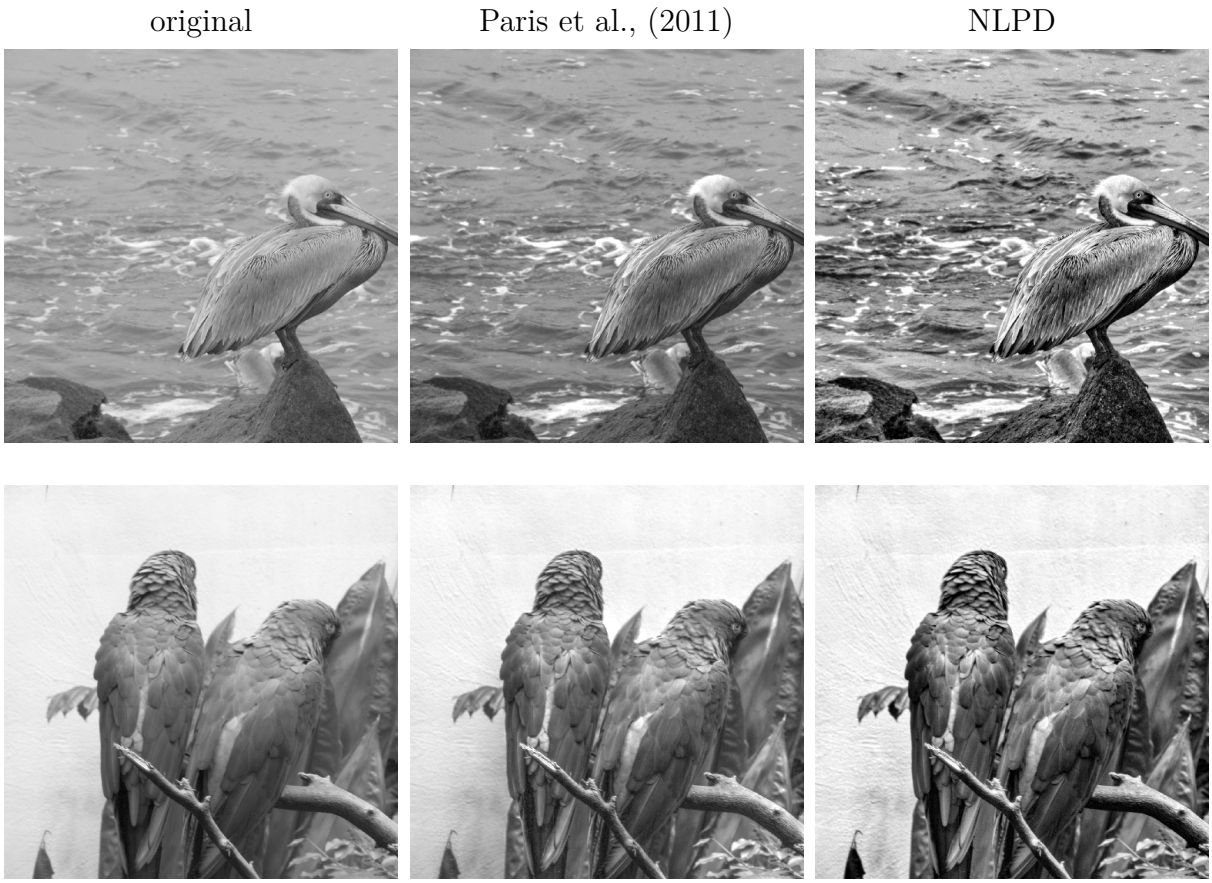

Figure 3.7: Rendering of two calibrated LDR images to a display with a limited luminance range of $[5, 300] cd/m^2$ (see caption of Fig. 3.6).

Rendering Calibrated HDR Luminances

We begin by considering rendering of images $\boldsymbol{S}$ obtained from a calibrated HDR imaging device, such that we know the true luminance values of all pixels. As an example, Fig. 3.6 shows an image from the database of Mark Fairchild (Fairchild, n.d.), with luminance range $S_{\min} = 0.78$ to $S_{\max} = 16200\, cd/m^2$. We wish to display this image on a device with a limited luminance range of $I_{\min} = 5$ to $I_{\max} = 300\, cd/m^2$ (typical values for a computer monitor). We solve for the perceptually optimal rendered image:

$$\hat{I}(\boldsymbol{S}) = \arg\min_{\boldsymbol{I}} D(\boldsymbol{S}, \boldsymbol{I}), \quad \text{s.t. } \forall i : I_{\min} \leq I_i \leq I_{\max}. \tag{3.19}$$

Figure 3.6 shows the original image intensities, linearly rescaled to fit within the luminance range $[I_{\min}, I_{\max}]$, an image tone-mapped using a recent state-of-the art method by Paris et. al. (Paris et al., 2011), and our perceptually optimized image $\hat{I}(\boldsymbol{S})$. The second image was computed using the default parameters recommended by the authors for tone mapping of HDR images: $\alpha = 1$, $\beta = 0$, and $\sigma = \log 2.5$. Linearly rescaling yields a rendered image in which most of the the details cannot be seen or differentiated. The algorithm by Paris et. al. (Paris et al., 2011) does an excellent job in mitigating this problem, rendering an image that reveals detail in both dark and bright regions. Nevertheless, the solution appears less detailed and lower in contrast than the image computed using our method. This is mostly because the Paris algorithm does not take into account the display luminance range. Although it (and most other tone-mapping algorithms) has additional parameters that can be adjusted, it is not obvious to a naive user how to select these parameters based on the display properties. In contrast, our solution is fully automatic (assuming the luminance values of the source image and the range of the display are

known), albeit at the expense of significantly more computation.

Rendering LDR Images with an Image Acquisition Model

Our method can also be used to improve the appearance of images acquired with a conventional low dynamic range (LDR) digital camera that has been calibrated to allow recovery of luminance values from recorded pixel values, $\boldsymbol{R}$. For most modern digital cameras, the acquisition luminance range is still generally much larger than the display range, and in any case, is unlikely to be exactly matched. Thus, we need to solve the following optimization problem analogous to the previous section:

$$\hat{I}(\boldsymbol{R}) = \arg\min_{\boldsymbol{I}} D(g(\boldsymbol{R}), \boldsymbol{I}), \quad \text{s.t. } \forall i : I_{\min} \leq I_i \leq I_{\max} \tag{3.20}$$

where g is the mapping from recorded pixel values to estimated scene luminances (in cd/m^2).

Results for two example grayscale images from the McGill database (Olmos & Kingdom, 2004) are shown in Fig. 3.7. For each image, we again compare the original image intensities, linearly rescaled to fit within the luminance range $[I_{\min}, I_{\max}]$, to our perceptually optimized image $\hat{I}(\boldsymbol{R})$, and a tone-mapped image computed using the Paris et. al method. (Paris et al., 2011). For the latter, we have again used the parameters recommended by the authors for tone mapping of HDR images: $\alpha = 1$, $\beta = 0$, and $\sigma = \log 2.5$. Our method again offers a visual advantage, producing higher contrast and more visible details. The improvement here is perhaps even more noticeable than in the HDR case, for which the Paris et. al. algorithm was developed.

Rendering Uncalibrated HDR Images

Unlike the situation in section 3.3.1, the typical scenario for images acquired from HDR cameras is that they are uncalibrated. That means that we have access to measurements $\boldsymbol{L}$ that are linearly related to actual luminances, but we do not have access to the scaling parameters (for instance, they might be normalized values, lying between 0 and 1). To apply our method, the measurements need to be linearly rescaled to luminance values, which amounts to estimating the minimum and the maximum luminance in the original scene ($S_{\min}$ and $S_{\max}$, respectively). One can often use an educated guess for those values given the content of the image – for instance, the luminance of a filament of a clear incandescent lamp is roughly $10^6\, cd/m^2$. As in the previous experiments, we solve the resulting optimization problem:

$$\hat{I}(\boldsymbol{S}) = \arg\min_{\boldsymbol{I}} D(\boldsymbol{S}, \boldsymbol{I}), \quad \text{s.t. } \forall i : I_{\min} \leq I_i \leq I_{\max} \tag{3.21}$$
$$\text{where } \boldsymbol{S} = (S_{\max} - S_{\min}) \cdot \boldsymbol{L} + S_{\min}.$$

Figure 3.8 shows the results for the widely-used HDR image "Memorial" for different values of $S_{\max}$ (and a fixed value of $S_{\min} = 5$). The proposed method converges to an image that exhibits enhanced contrast in all the regions, preserving the details, but also preserving the relative contrast and luminance between regions. This is particularly evident in high luminance regions (for instance the bright window behind the altar, or the round window in the top of the dome), where both the perceived details and luminance intensity is effectively portrayed.

As we increase the assumed maximum luminance of the original scene (while fixing the display restrictions), our algorithm further amplifies the contrast of details in the image.

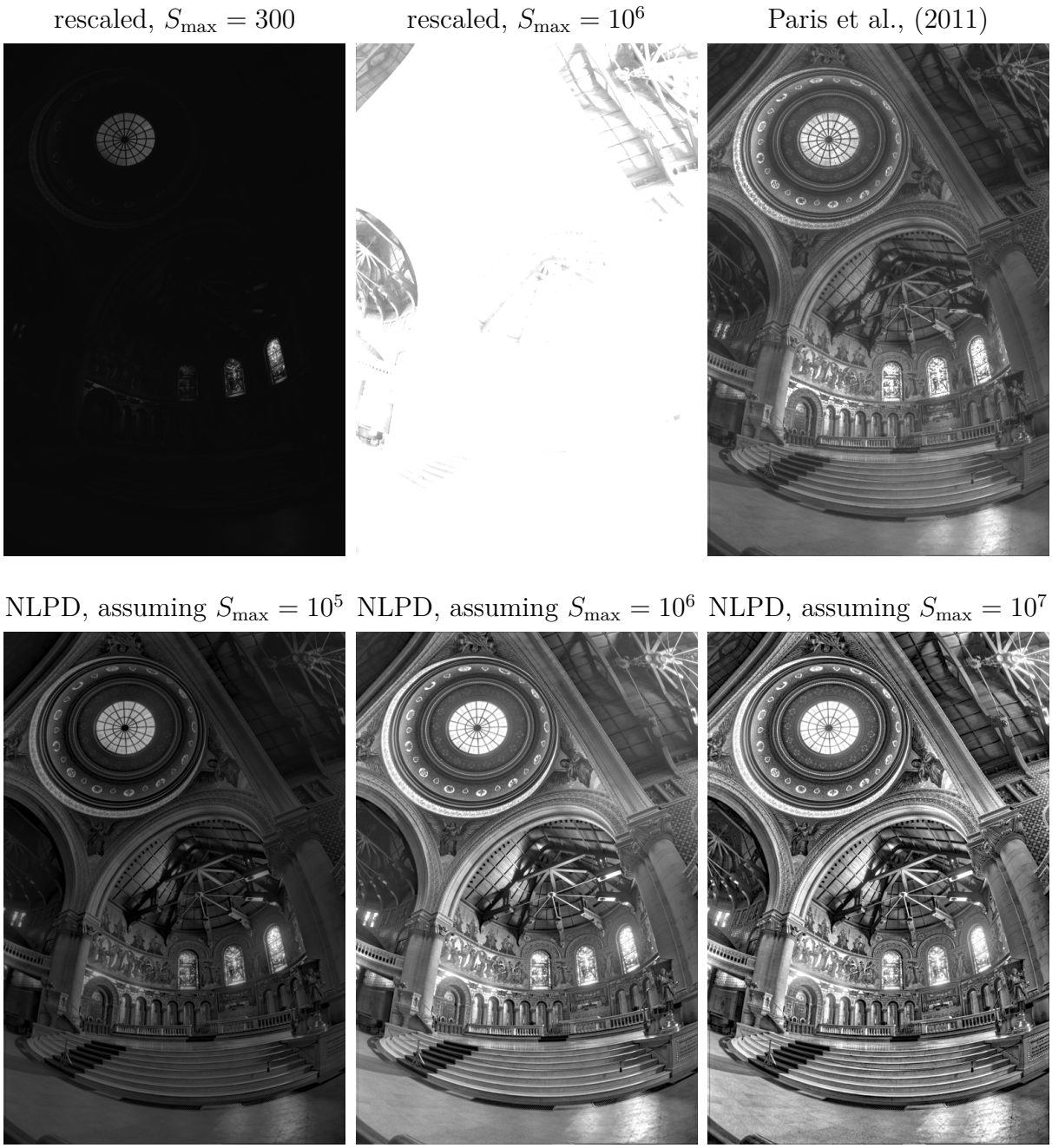

Figure 3.8: Rendering of an uncalibrated HDR image on a display with a limited luminance range. Linear mapping of luminances leads to loss of detail (top left: rescaling of luminances to the display range, assuming $S_{\max} = 300\, cd/m^2$; top center: rescaling of luminances, assuming a more realistic value of $S_{\max} = 10^6 cd/m^2$). Top right: the image rendered using (Paris et al., 2011). Bottom: the image optimized for NLPD, with different assumed maximum luminance values.

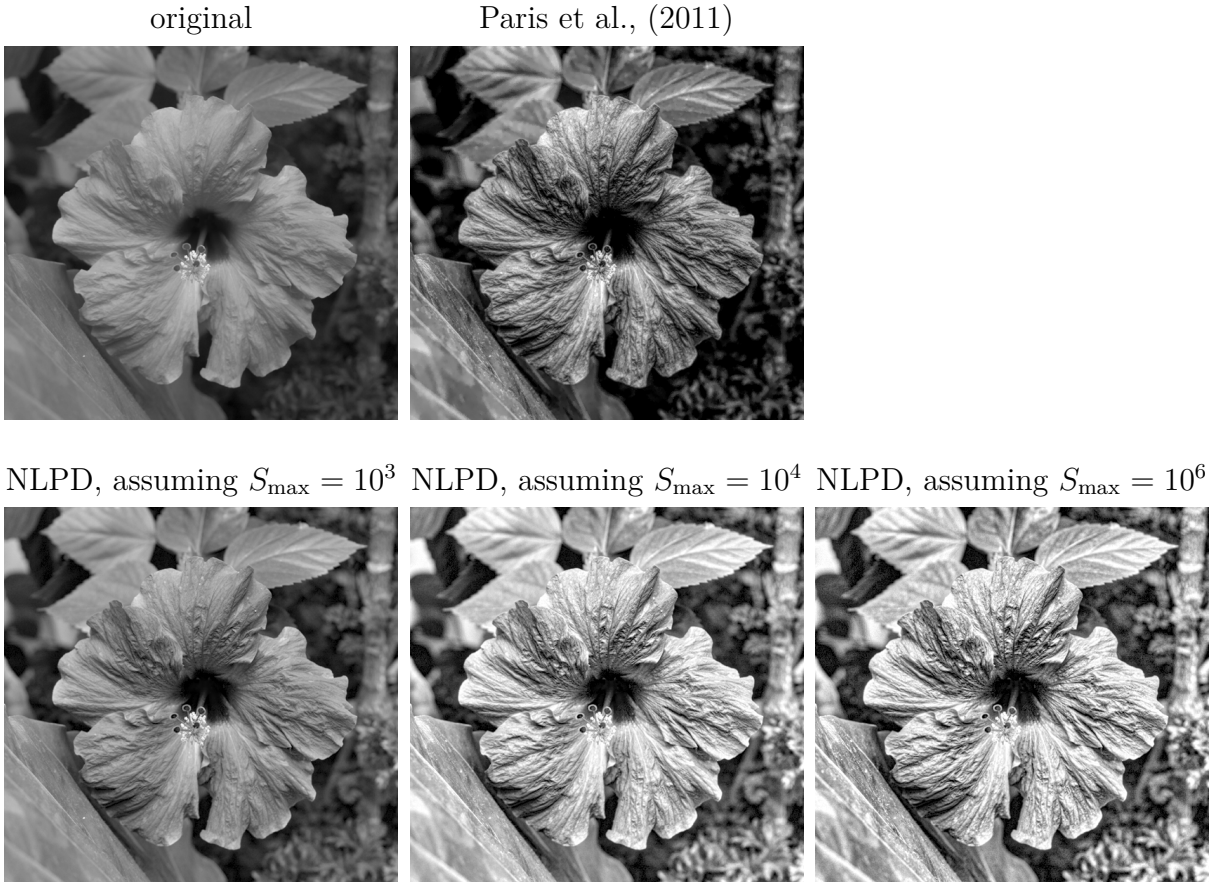

Figure 3.9: Example of artificial detail enhancement by simulating more light in the original scene. Top left: original image. Top center: image enhanced using (Paris et al., 2011). Bottom: image optimized for NLPD, with different assumed values of maximum luminance.

This makes sense from a perceptual perspective: If the original scene was brighter, an observer would be able to perceive more details within the scene. Therefore the method has to artificially enhance these details to mimic the appearance of the original scene. In the next two sections we take advantage of this behavior.

3.3.2 Artificial Detail Enhancement

We showed in the preceding sections that using knowledge about the image acquisition process helps greatly in automatically rendering images, given the display constraints. In some cases, however, detail visibility in the scene might be unsatisfactory. Intuitively, photographers know that the amount of detail visible in a scene depends on the amount of available light. If the image has already been acquired, it is of course not possible to alter the light sources. However, since the scene luminances scale linearly with the intensity of the light sources, our method allows us to simulate increased intensity post hoc, by linearly re-scaling the luminances of the scene, $\boldsymbol{S}$.

Figure 3.9 shows the results of modifying our choice of $S_{\max}$ (as in the previous experiment we fixed $S_{\min} = 5$). Note that with increasing values of $S_{\max}$, details become more visible. We also show the results of applying the Paris et al. algorithm, for which we have employed the parameters proposed in their paper for the detail enhancement problem: $\alpha = 0.25$, $\beta = 1$, and $\sigma = 0.3$.

3.3.3 Haze Removal

Surprisingly, this same method of detail enhancement can also be used for the problem of haze removal. In a hazy scene, the local contrast has effectively been reduced (roughly speaking, but adding a constant level of scattered light) which makes detail more difficult

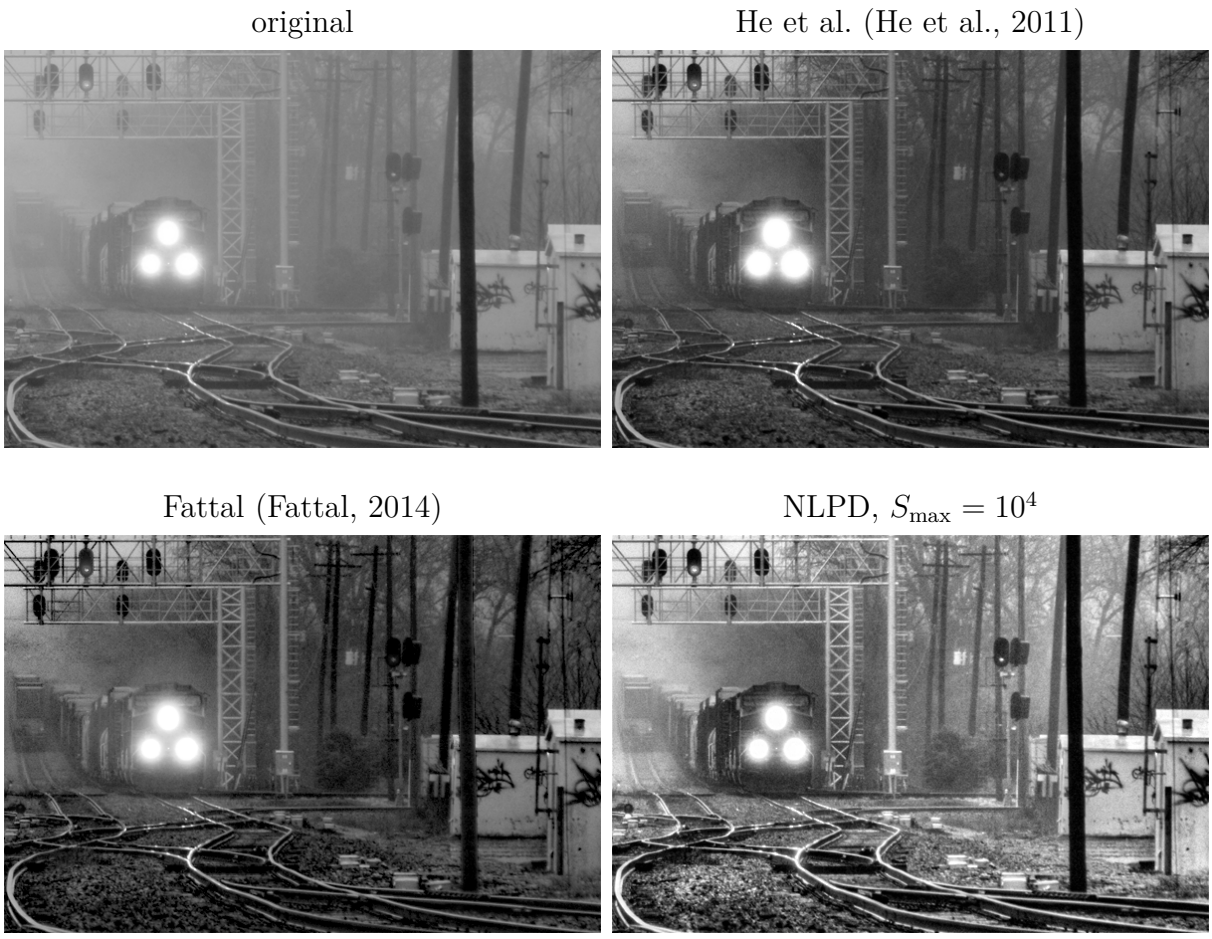

Figure 3.10: Example of haze removal. Top left: the original image. Top right: the image processed using He et al. algorithm (He et al., 2011). Bottom left: the image processed using Fattal algorithm (Fattal, 2014). Bottom right: the image processed by optimizing NLPD with $S_{\min} = 5$ and $S_{\max} = 10^4$.

to discern. In this experiment, we choose also $S_{\min} = 5$ (we find that results are fairly robust to the selection of this parameter) and $S_{\max} = 10^4$.

Figure 3.10 compares the performance of our method with two other methods (Fattal, 2014; He et al., 2011). Our algorithm converges on an image that greatly enhances the details of the original hazy image, boosting the contrast and reducing the perception of haze within the image. Although the other two methods are specifically designed for this particular problem, our method obtains a similar result without modification.

3.3.4 Varying Display Constraints

While examining the effects of various image acquisition scenarios in the previous section, we assumed only that the display luminance is bounded. The upper bound is a natural constraint for any real display. The lower bound is also relevant for a wide range of practical display devices, and arises from reflected ambient light and scatter within the display. In this section, we examine the effect of each of these constraints independently, along with a few more complex constraints.

Figure 3.11 shows the results for different minimum and maximum luminance bounds, ($I_{\max}$, $I_{\min}$). Our method enhances local contrast, whereas linear rescaling can only manipulate contrast globally. For a wide range of display characteristics, optimizing the image to minimize the NLP distance reduces distortion in the rendered images, and increases the visibility of perceptually relevant features.

Rendering with Limited Power Consumption

The proposed framework allows us to seamlessly introduce arbitrary display constraints. For example, we can optimize the trade-off between image quality and power consumption.

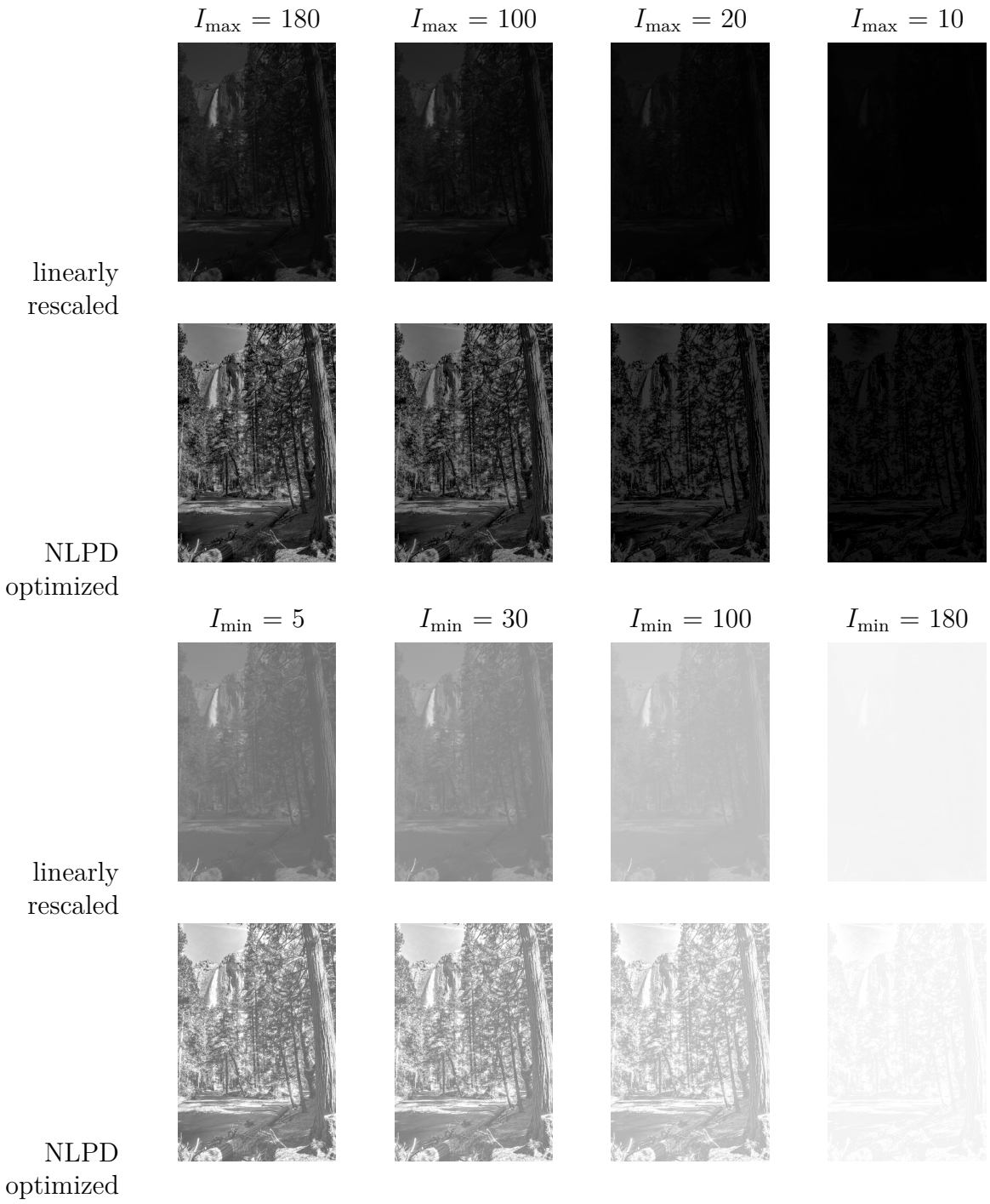

Figure 3.11: Effect of different maximum and minimum display luminance constraints. Top two rows: The image rendered for different levels of maximum luminance (assuming $I_{\min} = 5$), by linearly rescaling (1st row) versus NLPD-optimization method (2nd row). Bottom two rows: analogous, but for different levels of minimum luminance (assuming $I_{\max} = 300$).

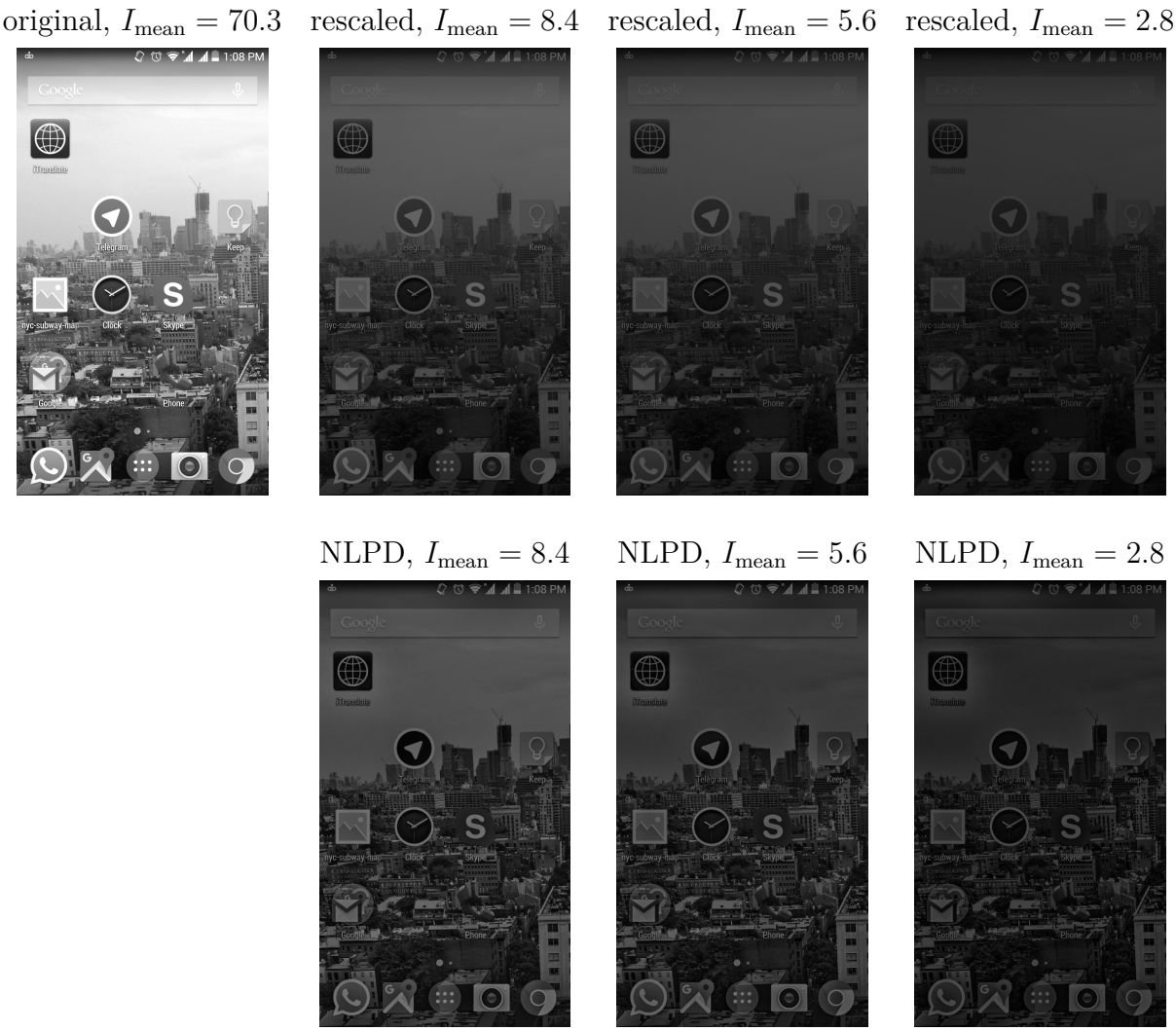

Figure 3.12: Rendering with a power consumption constraint. Top left: the image at full luminance (smartphone screenshot). Top row: the image linearly rescaled to achieve target mean luminance. Bottom row: the image optimized for NLPD with target mean luminance constraint. Assuming power consumption is proportional to mean luminance, the NLPD-optimized renderings convey more detail than their linearly-rescaled counterparts, while consuming the same power.

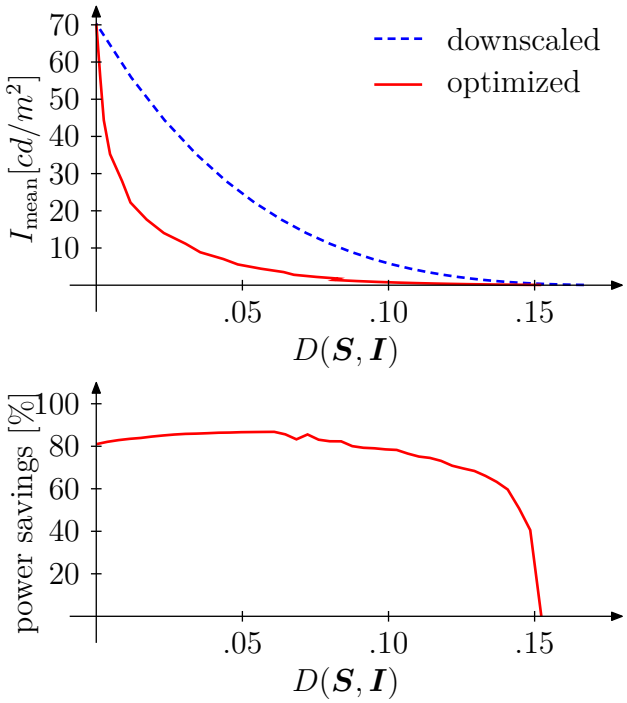

Figure 3.13: Trade-off between power consumption and image quality, comparing linear luminance rescaling to optimization of perceptual distortion with a mean luminance constraint. Top: the relationship between perceptual distortion $D(\boldsymbol{S}, \boldsymbol{I})$ and mean display luminance I_{mean}. For any given acceptable distortion level, the optimization method requires only a fraction of the display luminance, hence significantly decreasing power consumption. Bottom: power savings, quantified as one minus the ratio of required mean display luminances for the two methods.

To illustrate this, we assume power consumption is proportional to mean display luminance (as for instance in organic light-emitting diode displays used in cell phones - if the relationship were nonlinear, that could also be incorporated into the problem), and optimize the NLPD while constraining both the mean luminance as well as the range:

$$\hat{I}(\boldsymbol{S}) = \arg\min_{\boldsymbol{I}} D(\boldsymbol{S}, \boldsymbol{I}), \quad \text{s.t. } \forall i : I_{\min} \leq I_i \leq I_{\max} \text{ and } \frac{1}{N_i}\sum_i I_i = I_{\text{mean}} \tag{3.22}$$

Figure 3.12 shows images optimized for different mean luminance values compared to images linearly rescaled to achieve the same target mean luminance. For each mean luminance value, the NLPD-optimized images retain more detail from the original scene than the rescaled images. In figure 3.13, we plot mean luminance as a function of perceptual distortion (NLPD) for both methods. Optimizing the images yields a clear benefit in terms of the trade-off between mean luminance and perceptual distortion. Over a wide range of distortion levels, we see that the NLPD-optimized images reduce power consumption by roughly 80% compared to linear rescaling.

Rendering with a Discrete Set of Gray Levels (Dithering)

Most displays have a limited number of available gray levels. In the extreme case this can be as few as two (e.g., black-and-white printers, e-ink devices, etc). Here, we illustrate that the proposed method is flexible enough to produce good results even under such extreme constraints. The optimization problem is the same as before, but here, we restrict the pixel values to be taken from a discrete set:

$$\hat{I}(\boldsymbol{S}) = \arg\min_{\boldsymbol{I}} D(\boldsymbol{S}, \boldsymbol{I}), \quad \text{s.t. } \forall i : I_i \in \{I_{\min}, \ldots, I_{\max}\}. \tag{3.23}$$

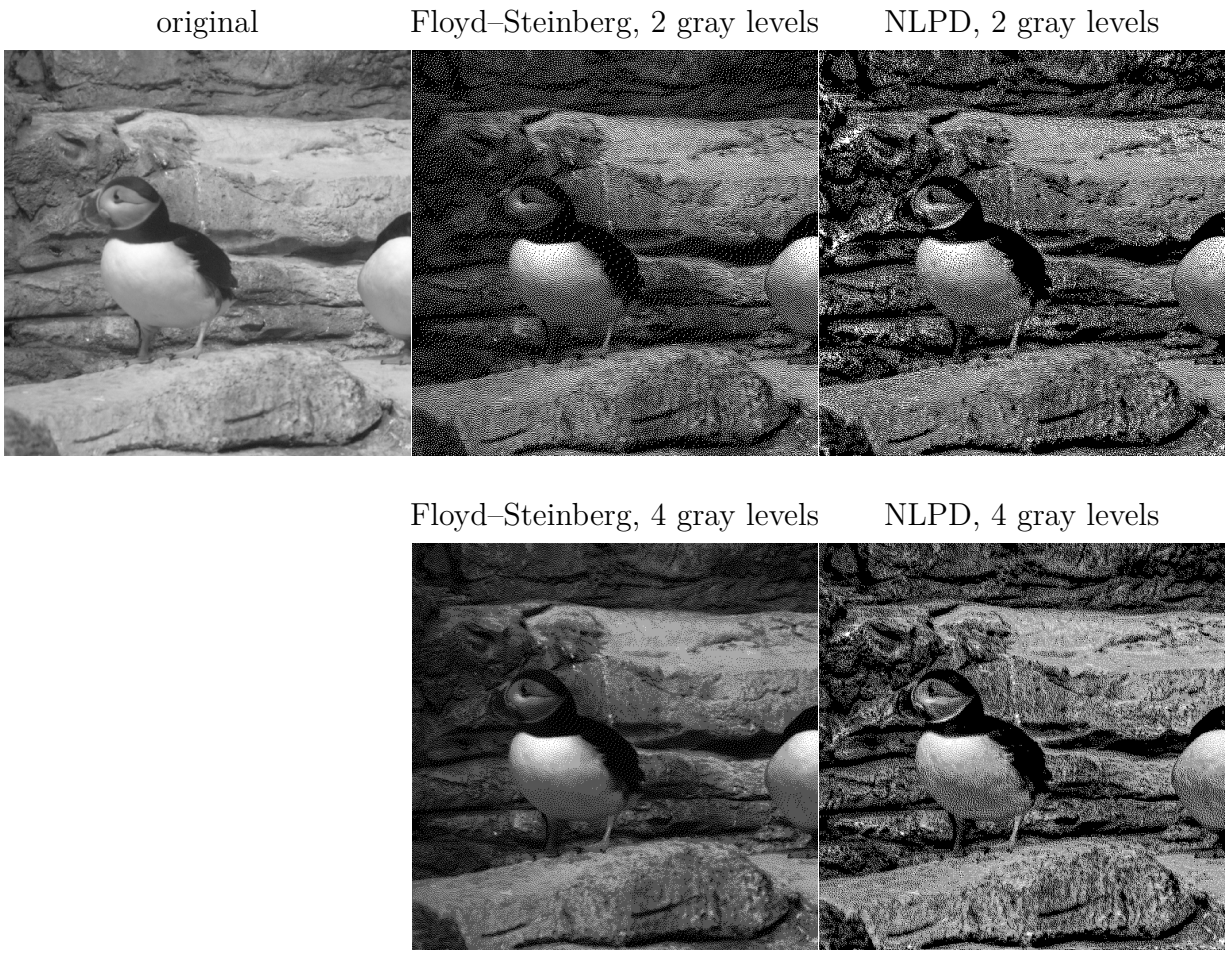

Figure 3.14: Rendering with a discrete set of gray levels. Top left: the original image. Center column: the image rendered with two or four gray levels using a standard error diffusion (Floyd–Steinberg) method (Floyd & Steinberg, 1976). Right column: the image rendered with NLPD error diffusion.

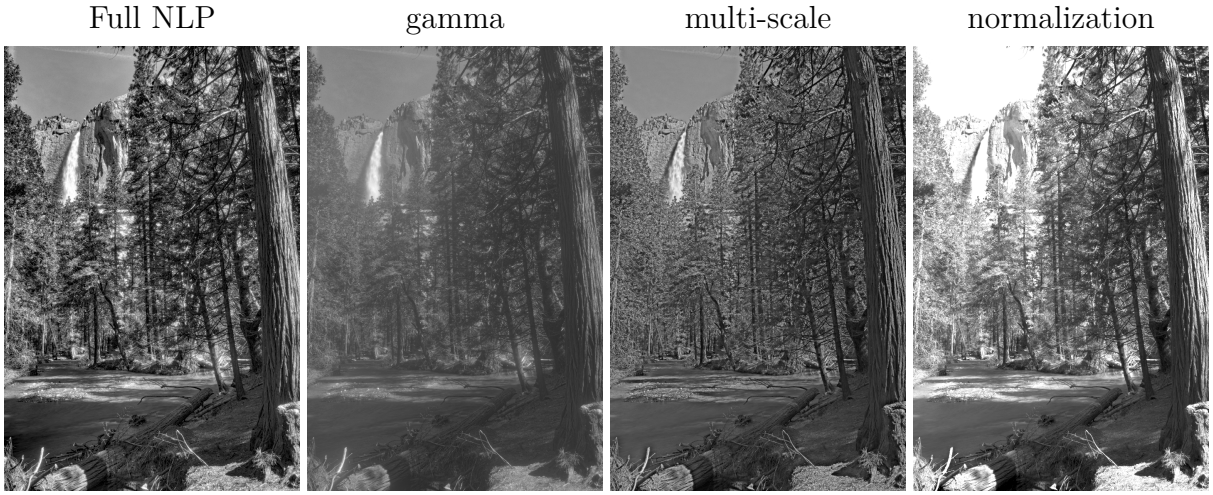

Figure 3.15: Rendering of an HDR image with different parts of the NLP transformation removed (see text). The ablated component of the model is listed above each example.

The discrete nature of the optimization problem prevents us from using a gradient-based method. Instead, we use a greedy error-diffusion algorithm, analogous to the classic Floyd–Steinberg method. We first initialize the image to the solution obtained for a continuous range of luminances, as in previous experiments. Then, we iteratively select the discrete value for each pixel of the image in raster-scan order, each time picking the discrete value that minimizes the NLP distance of the intermediate result to the original scene.

Figure 3.14 shows the results for images rendered using two and four gray levels. In low contrast regions, our method is seen to preserve significantly more detail than the Floyd–Steinberg method. In addition, the Floyd–Steinberg algorithm tends to generate artificial patterns in extensive regions of slowly-varying luminance, which can be seen in the dark regions of the bird's wings. Our method, however, does not generate these artificial patterns.

3.3.5 Contribution of Perceptual Metric Components

To provide intuition regarding the effect of each of the primary components of the NLP, we optimized images for rendering while removing one of three components of the transform: the initial point-wise nonlinearity (set $\gamma = 1$), the multi-scale decomposition (set $N_k = 1$), and divisive normalization (set $P = 0$ and $\sigma = 1$). Figure 3.15 shows results for each manipulation. Note that we did not refit each of the partial transforms to predict human perceptual judgments; therefore, these results should be seen as a way to understand the importance of each computation, and not as a quantitative comparison of image quality assessment performance (see details in Appendix C).

Each of the three images differs noticeably from the one optimized with the full transform. Without the initial point-wise nonlinearity, the algorithm produces images in which low to medium luminance patches of an image are misrepresented. The high luminance areas are detailed but some parts with medium or low luminance are reduced in contrast. Without the multi-scale decomposition, the algorithm produces images in which extremely high and extremely low frequencies are well preserved, but intermediate frequencies are underrepresented, and in some cases nearly disappear. And without the contrast normalization, the algorithm converges to images that saturate at the luminance boundary constraints of the display. Normalization preserves the relative luminance changes between coefficients while allowing the absolute luminance to be modified. This allows the rendered image pixel intensities to be proportional to the relative power in each local region. Moreover, this ensures that regions with similar content scale in a similar way.

3.4 Summary and Extensions

We have described a framework for directly optimizing rendered images, taking into account display limitations, so as to minimize perceptual differences between the rendered image and the original scene. The method is parameter-free and only requires knowledge of the display restrictions and the original scene intensities. Since these restrictions are expressed in standard physical units (cd/m^2), if either is missing, suitable values can be estimated easily. We have shown that our method matches or exceeds the state-of-the-art for rendering across a variety of acquisition conditions and display restrictions.

We've employed a perceptual metric based on an abstraction of the transformations implemented in the early stages of the human visual system. The metric is an extension of the NLP distance presented in (Laparra et al., 2016), adapted to deal directly with luminances and images of any size. We fit the parameters of this metric to optimize its ability to predict human distortion ratings. We have shown that this metric is consistent with human perception, exhibiting correlation with human quality ratings that is similar to or better than full-reference models specifically designed to assess perceptual quality (see appendix C). It is continuous and has well-behaved gradients, making it easy to incorporate into a rendering optimization framework. In addition, it has also been previously employed to optimize an image compression algorithm (Ballé et al., 2016).

Most contemporary tone mapping methods do not make explicit use of perceptual metrics (see Cerdá-Company et al., (2016) for a nice review), but rather provide the user with a small set of free parameters to hand-adjust the mapping from scene to displayed image. These methods are conceptually simpler than ours, and some of them can produce high quality results in controlled situations (see for instance Paris et al., (2011)). Nevertheless,

their parameters are often difficult to interpret (and thus, to set), and the restriction to particular functional forms may limit their applicability to specific rendering problems.

In contrast, by directly optimizing the rendered image itself, our method is able to take into account different display constraints, without requiring manual selection of an appropriate parametric mapping for each situation, and without requiring a human operator to adjust any parameters. The downside of this approach is computational cost: optimization over the high-dimensional space of feasible rendered images is expensive, and although both hardware and software continue to improve, this optimization will always be significantly more expensive than optimizing a small set of parameters for a fixed transformation. Even if the computational costs prevent the use of this method in a real-world application, the results can still serve as a benchmark for what is possible, thus facilitating the development of alternative methods.

Although our framework may be applied to any display problem, the solution can depend heavily on both the perceptual metric employed, and the method used to solve the constrained optimization (for example, if the constraints force the problem into nonconvex or discrete regimes). Optimization has undergone dramatic changes in the past decade, and methods for handling nonconvex and discrete problems have become more reliable and efficient. As an example, we believe it will be possible to improve on our halftoning solution (for which we used a simple greedy method with error diffusion).

Our use of a simple physiologically-inspired model for assessing perceptual distortion also offers opportunities for improvement (note that most image quality models are less physiologically motivated (Narwaria et al., 2015; Wang et al., 2004, 2003)). For example, the NLPD can likely be improved by including relationships between frequency channels, which could help to control artifacts such as halos that sometimes appear around high-

contrast edges. In addition, the NLP model should be extended to operate on color images, and to include another stage of processing corresponding to primary visual cortex (for example, using oriented, multi-scale, derivative filters with cross-scale and cross-orientation normalization). All of these improvements can be made following the same framework that we have presented for the current model: defining a functional form based on the transformations of sensory neurobiology, fitting the parameters using human perceptual data, and using this model with fixed parameters to optimize the rendering of images.

3.4.1 Optimized Image Rendering as a Test of Perceptual Metric Quality

The rendering framework we developed relies critically on an accurate metric quantifying the perceptual differences between the rendered image and the original scene, and optimally-rendered images can thus provide a strong indication of the abilities of such a metric. Here, we use our optimal rendering framework to test different layers of VGG16 as perceptual metrics and compare them to the performance of our NLPD metric show above.

We evaluated the ability of the representations at the 6 layers of VGG16 analyzed in chapter 2 to serve as human perceptual metrics within our framework. Perceptual distance, $D(S, I)$, for each layer was computed as the Euclidean distance between that layer's representation of the scene, $f(S)$, and the representation of the rendered image, $f(I)$.

$$D(S, I)_f = ||f(S) - f(I)||_2 \tag{3.24}$$

We compare the results to the NLP results above, as well as two point-wise tone-mapping algorithms (linear-rescaling and non-linear "gamma" rescaling) that do not operate within our rendering framework. In the NL-Rescaled example, we take the 6th root of each pixel intensity before linearly rescaling the image to fit within the displayable range.

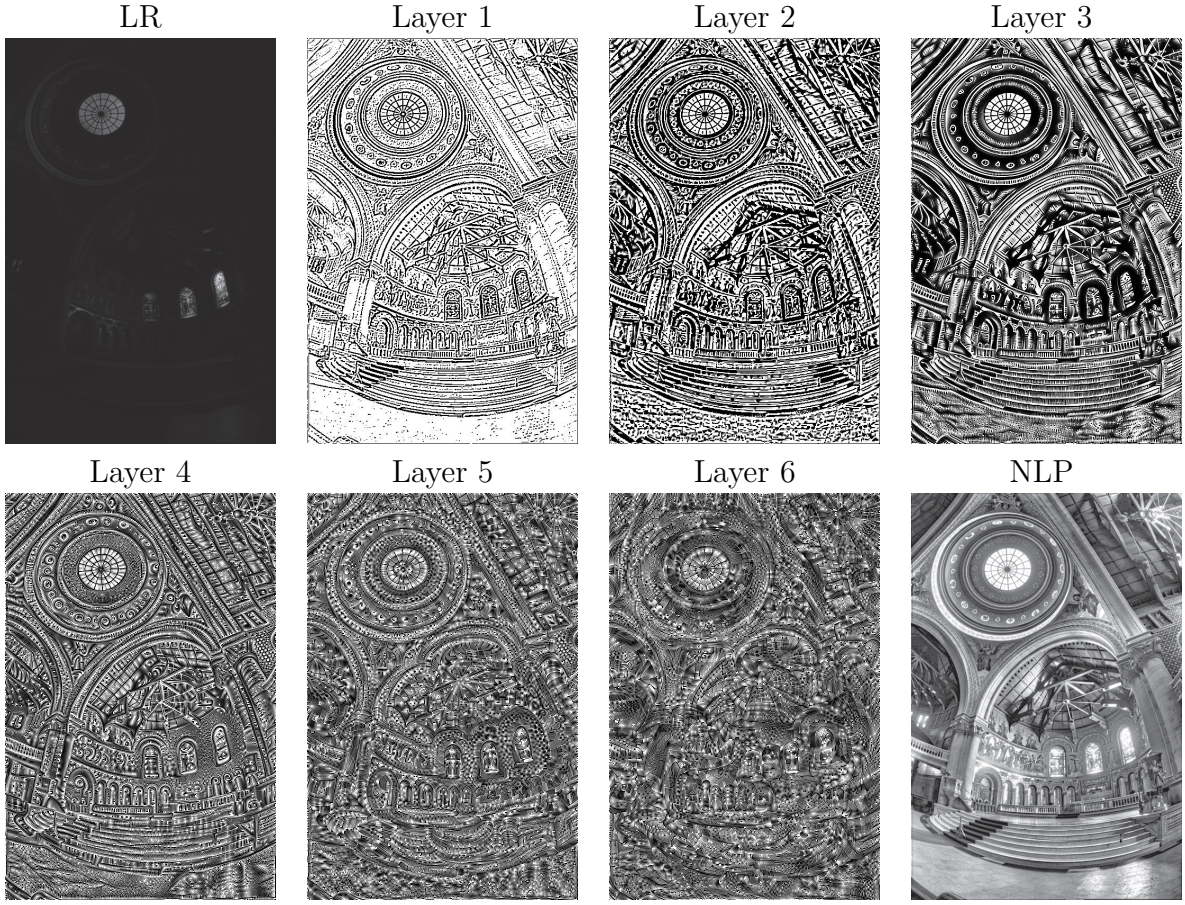

Figure 3.16: Images optimized using different perceptual models (see text). **Original Scene Luminance**: Minimum: $0\frac{cd}{m^2}$, Maximum: $10^7\frac{cd}{m^2}$. **Displayable Luminance**: Minimum: $5\frac{cd}{m^2}$, Maximum: $300\frac{cd}{m^2}$.

Figure 3.16 shows the results for the optimization of the "Memorial" HDR image, under the same display constraints, for each of our candidate perceptual metrics. Consistent with results from chapter 2, we see that early layers (1-3) of VGG16 are better perceptual metrics than Pixel MSE, and also better than deeper layers (4-6). The early layers generate images that capture near-binary renditions of distinct image features, but discard nuances of lighting, reflectance and shading. Later layers replace correct image content with hallu-

cinated artifactual features (e.g., swirls). The LGN model, on the other hand, generates a natural looking image, successfully balancing the content and contrast of the rendered image better than any of the VGG16 layers, and better than the linear or nonlinear tone-mapping solutions. Similary to our observer analysis in chapter 2, we now treat each model as an observer and ask them to evaluate the perceptual distance between the original scene and each fully-rendered image. Distances according to each observer model are reported below (See Figure 3.17). Each observer reports the shortest distance for its own synthesized image, and otherwise have very different predictions about the perceptual distance between each of the other image pairs. We ran a simple experiment that captured the rank ordering of the naturalness of the synthesized images according to human observers (n of 2), and compared the results to the rank order produced by each observer model. As expected from visual inspection, the NLPD model not only synthesized the most realistic image, but NLP distances also predict human rankings at a high level, higher than all layers of VGG16. The only model that improves upon NLP's predictions is our single-scale On-Off model tested in Chapter 2, which perfectly predicts the human rank order. No layer of VGG16 predicts the human rank order of these images at a high level.

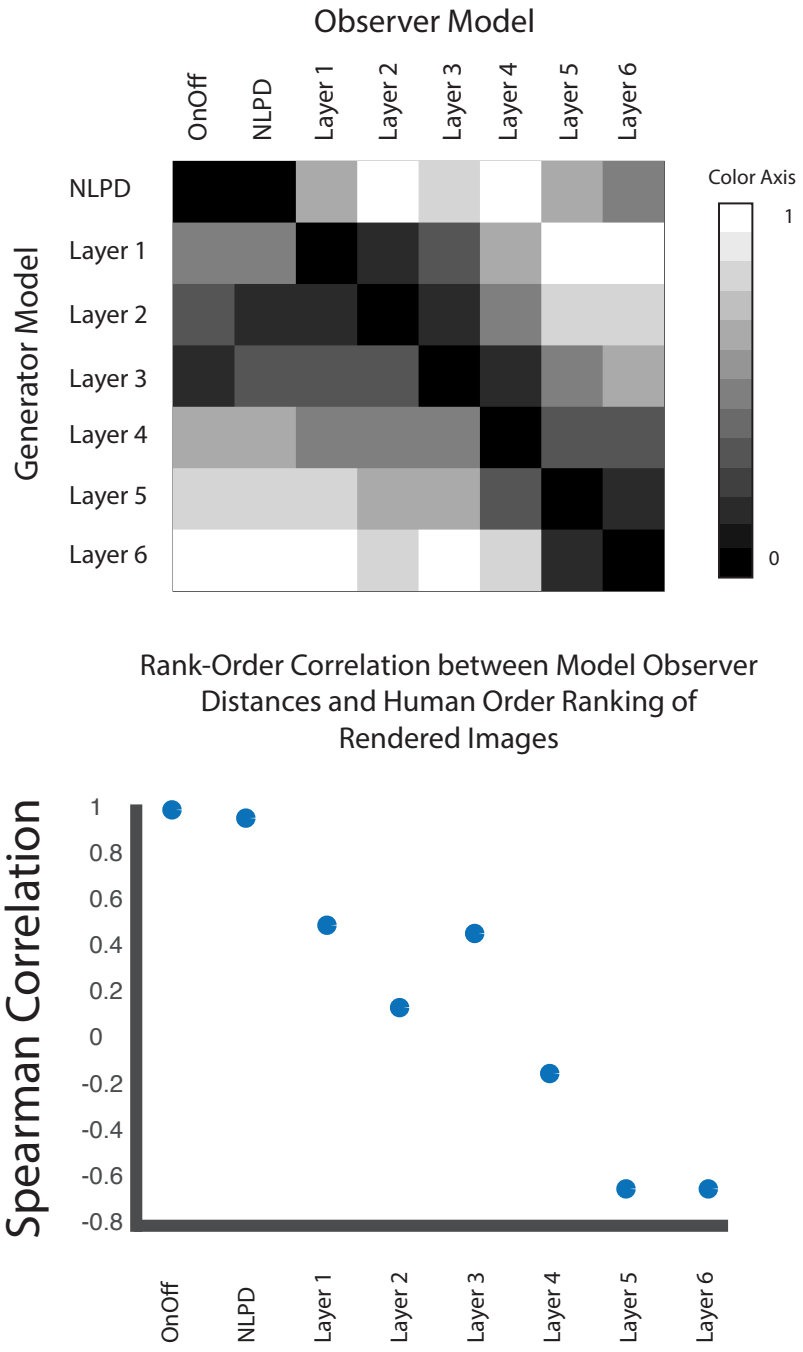

Figure 3.17: Observer Model's ranking of Rendered Images. **Top:** Rendering distances for each Observer Model (Columns) for images rendered using each Generator Model (Rows). **Bottom:** Observer model's Spearman (rank-order) correlation to human rankings of the "naturalness" of rendered images. Only the OnOff and NLP metrics perform well at predicting human rankings.

Chapter 4

Towards a Normative Model of Perceptual Distortion Sensitivity

In the preceding chapters, we attempted to capture human perceptual distortion sensitivity within various models inspired, to different degrees, by the architecture of the human visual system. We showed that models of this nature, fit to a database of human perceptual judgments, outperformed the state of the art in the field significantly by traditional measures of success at the task. In addition, we found that several different neural network inspired models performed equivalently well when measured by these traditional measures. These measures, while standard, fail to capture the entire extent of the space of possible image distortions. While they may be adequate if one wishes only to use the model to evaluate known forms of distortion encountered regularly in the transmission of images, for example errors introduced by a compression algorithm such as JPEG2000, performance on these measures may be adequate. However, many modern applications utilize these models to compute "perceptual loss functions", or objectives to be minimized in an optimization procedure, standing in for an average human observer. The nature of possible distortions encountered during an optimization of this kind covers much more of the poten-

tial distortion space than simple distortions included in an engineer's database. In order to differentiate the performance of these models in this high-dimensional space, and to gain insight into how each of them generalize to unseen types of distortions, we developed a model-constrained synthesis method for generating targeted test-stimuli that can be used to compare model sensitivity to human sensitivity.

Utilizing Fisher Information to predict model sensitivity to local perturbations of an image, we found each model's prediction of the most and least noticeable changes we could make to an image, and a corresponding prediction of how sensitive the humans should be to these changes. We compared these predictions to empirical human sensitivity to these changes. We found that, despite the fact that all of our neural networks explained data within our testing database equally well, they did not generalize equally well outside of the database. In fact, we found that the networks that were more closely based on known physiology, even the simple nonlinear computations of early visual physiology, generalized significantly better than any of the other neural networks.

In a parallel line of work, we generalized our model based on the LGN to operate at multiple scales, and found that models constructed to reflect the early visual system significantly reduce mutual information between model coefficients and their neighbors compared to image pixels and their neighbors. Creating a multi-scale representation is a necessary step for converting a model of the visual system optimized for a particular viewing distance into a model that can operate on images in the real world, where viewers may view any image from many different distances. In parallel, we developed a framework for optimally rendering images on a screen when the displayable luminance range of the screen is smaller than the range of luminance captured by the camera sensor. We showed that our multi-scale LGN model (NLPD) outperformed all of the other models we tested

at explaining human perceptual data, including our own single scale LGN models. We then put this model to work as "perceptual loss function" in our optimized image rendering framework. We showed that this model performed well as a perceptual loss function in our rendering framework, and in addition, showed that it performed significantly better as a perceptual loss function than the other neural networks, trained on object recognition, that we had tested in previous analyses. This result buttressed our previous results, showing that neural networks that include more known physiology generalize better in the high dimensional space of potential image distortions introduced by real-world applications.

While our models of early visual physiology seem to describe human perceptual sensitivity better than other models, there are certainly dimensions along which such a simple model is unable to distinguish true distortions from natural changes to images (such as changes to semantic scene and object identity information). In addition, we do not have a ground truth measure of how close the models are to the true model of human sensitivity. It's quite likely that we can create image perturbations that are both more and less noticeable than even the very good predictions that come from our best performing model, but it is difficult to know how to achieve this without already having a perfect model of human vision. In the absence of this perfect model, it may be possible to draw insight from a normative model that describes the types of real-world image transformations that humans are sensitive and insensitive to. It should also be possible to build this sort of knowledge into a model of perceptual similarity. In fact, Zhou Wang and Eero Simoncelli explored this line of thinking in their conference paper "An Adaptive Linear System Framework for Image Distortion Analysis" (Wang & Simoncelli, (2005)). There are two key insights in their paper. The first is that image distortions could be decomposed into "structural" distortions, those that change the structure or identity of the objects in the scene, and

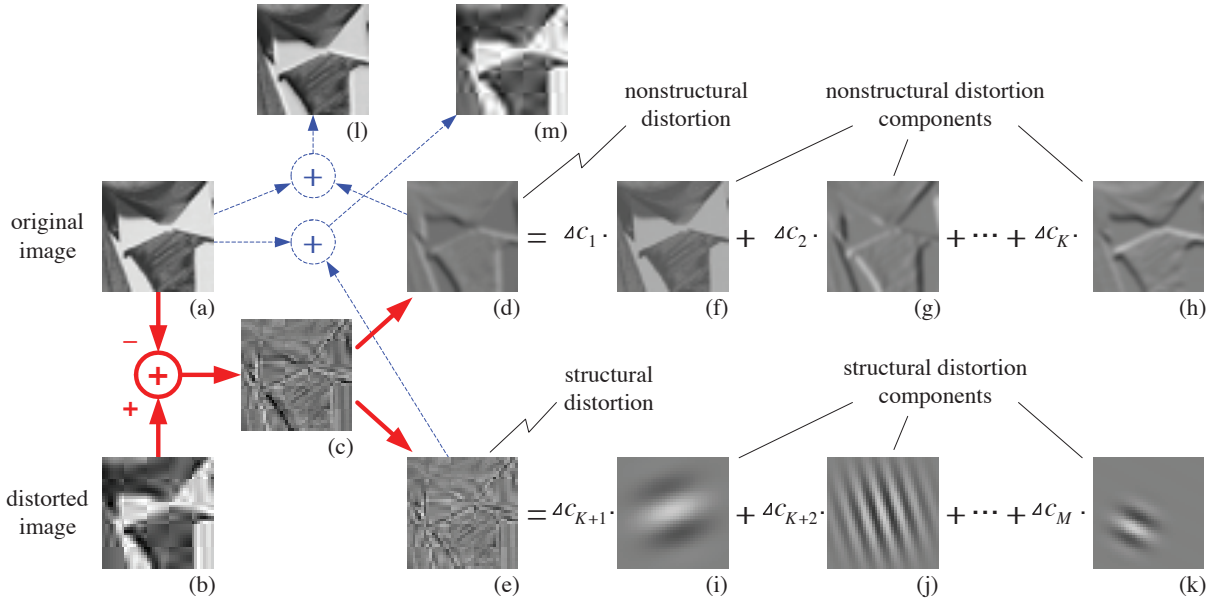

Figure 4.1: Separation of structural and non-structural distortions using an adaptive linear system (Adapted from Wang & Simoncelli, (2005))

"non-structural" distortions, those that don't. The authors hypothesized that the human visual system is built to be much more sensitive to the former than the latter. The second insight is that non-structural components can not be computed from a fixed basis of linear filters, but must be adaptively computed from the input signals themselves. In that work, they constructed a metric from the combination of the discrete cosine transform basis (structural components), and a small set of local Taylor expansions along different transformation dimensions (non-structural), and showed that this metric explained human distortion judgments well (Wang & Simoncelli, (2005)).

Both of these insights help explain why our more nonlinear models perform better than simpler linear and quasi-linear models of neural processing, like the LN model examined here, or other carefully constructed linear models such as (See figure 4.2 for an example) (Watson, (Jan 2000) and Watson & Ahumada, (2005)). The simpler linear models operate

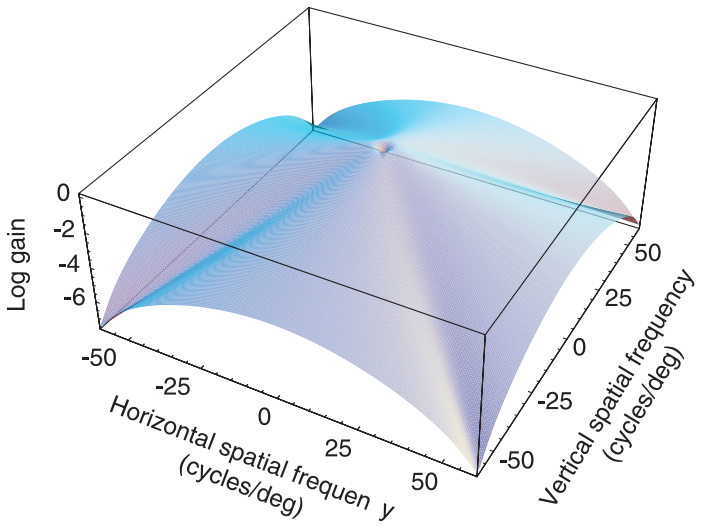

Figure 4.2: A well characterized model with non-adaptive structural (spatial-frequency) sensitivity. While this model carefully captures human sensitivity to spatial frequencies (as measured on controlled stimuli), it makes the same set of predictions for the most and least-noticeable change for every image (noise composed of the spatial frequencies found at its peak and trough, respectively) regardless of the underlying content. (Adapted from Watson & Ahumada, (2005))

on a fixed (or mostly-fixed) basis, our models have components within them (the divisive normalization modules), which are not fixed, but adapt to the input signal itself. We can also map the eigen-distortions produced by our best performing models, such as OnOff, onto this framework. On-Off's predictions map loosely to a combination of structural and non-structural changes, with a slight modification. The Most-Noticeable eigen-distortion from a model with non-adaptive structural components, such as the LN model, is noise composed of the spatial frequency that the model is most sensitive to dispersed across the image and is equivalent regardless of the underlying image. The same is true for its least-noticeable prediction, however it is now noise composed of the spatial frequency that the model is least sensitive to (See Figures 4.2, 4.3, and 4.4).

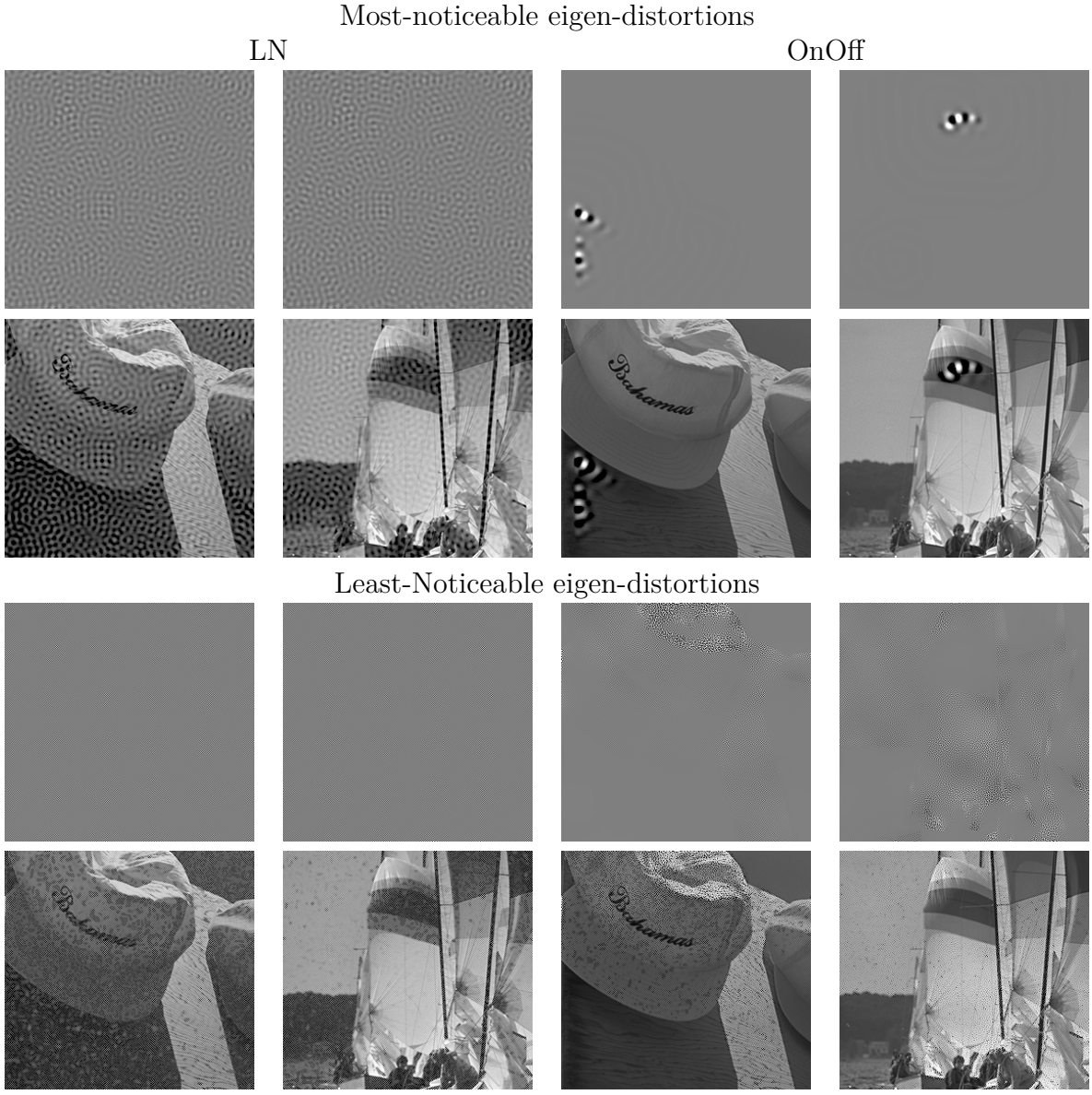

Figure 4.3: **Comparing Predictions From Non-adaptive and Adaptive Structural Representations.** Predictions from the non-adaptive LN model are the same regardless of underlying image content. Predictions from the Adaptive OnOff model change spatial frequency content based on the underlying image content.

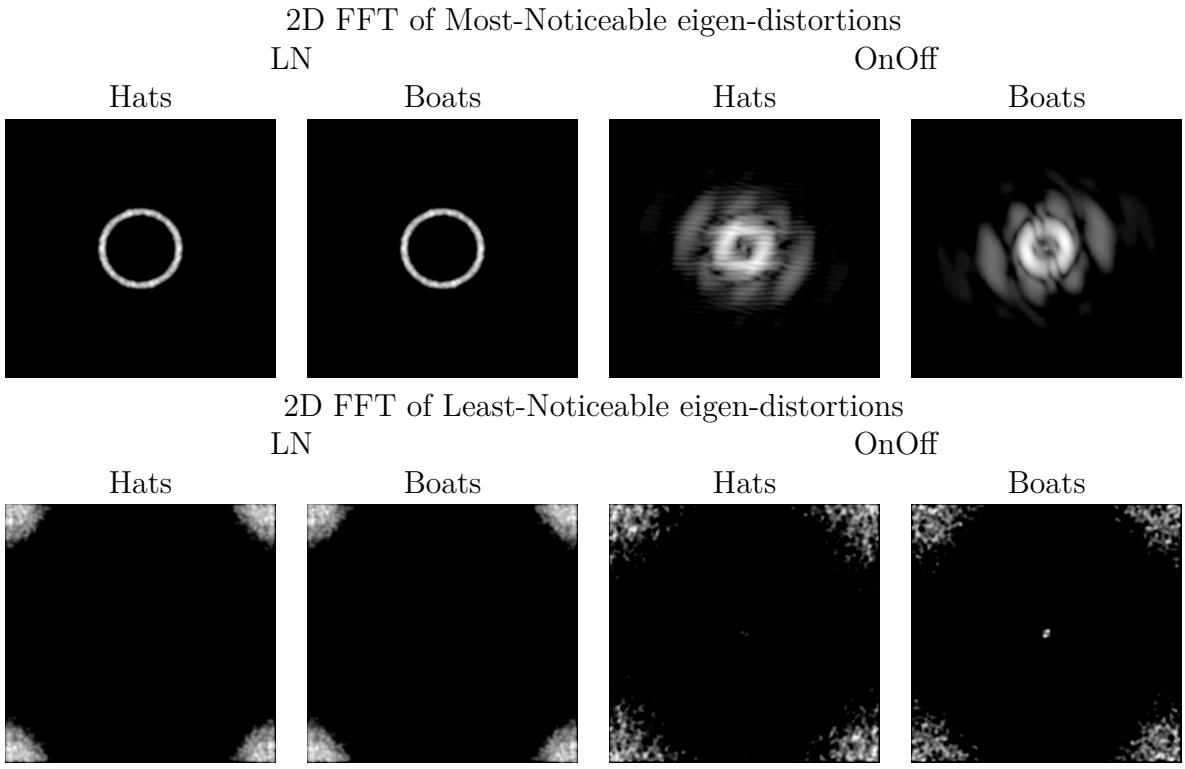

Figure 4.4: **Comparing Spatial-Frequency Content of Eigen-distortions from Non-adaptive and Adaptive Structural Representations.** The predictions for two different images (Hats and Boats), from the non-adaptive LN model have the same 2-D Fourier spectra, indicating that the model is image agnostic. The predictions for the OnOff model, however, have very different 2-D Fourier spectra, indicating that the OnOff model adaptively changes its sensitivity based on underlying image content.

The OnOff predictions, however, are local, image specific, patches of the model's most sensitive (or least sensitive) spatial frequencies, because the OnOff model adapts its sensitivity to structural distortions based on local image content as well as its sensitivity to non-structural changes, such as luminance and contrast changes (See Figure 4.3 and Figure 4.4). This ability to adapt its sensitivity simultaneously to structural and non-structural distortion components based on the content of the image is what makes the OnOff model so powerful.

Though Wang and Simoncelli tested a perceptual sensitivity metricconstructed based on these principles, they did not test this hypothesis directly using psychophysics, because of many real-world limitations. However, the advent of high quality, physics-based image rendering tools allows us to overcome those barriers and take steps towards testing this hypothesis directly. The rendering environment gives us the ability to create images modified by independently modifying elements that define scene elements, such as object location and orientation, as well as image capture elements, such as the 3 dimensional location of the camera within the scene. The work presented below is the start of an effort to do just that, in collaboration with David Brainard at the University of Pennsylvania. For this experiment, we utilize Professor Brainard's Virtual World Toolbox, built on RenderToolbox4, to render images that differ only along several non-structural dimensions defined by rendering parameters, and compare human sensitivity along each of these dimensions (Heasly et al., 2014). We also compare human sensitivity along these dimensions to human sensitivity to our OnOff models eigen-distortions (representing adaptive structural and non-structural changes to the images). Finally, we analyze how well the OnOff model captures human sensitivity along each of these dimensions in order to analyze potential directions for improvement. This work is in progress and as such represents an incomplete picture.

4.1 Developing a Normative Model

In their paper, Wang and Simoncelli defined non-structural perturbations as "gentle distortions caused by variations of lighting conditions, spatial movement, or pointwise monotonic intensity changes caused by image acquisition and display devices that should not change the perceived structure" (Wang & Simoncelli, (2005)). In that work, they defined these

axes as luminance changes, contrast changes, gamma distortion (a pointwise nonlinear intensity distortion caused by image acquisition and display devices), as well as horizontal and vertical translation (Wang & Simoncelli, (2005)). The structural distortions are what is left over after removal of the non-structural components. Here, we attempt to modify our base images along axes that roughly correspond to these axes. We also draw a distinction between translation and motion of objects within and scene, and translation and motion of the camera, which causes coherent changes across the entire scene at once.

4.1.1 Distorting Images Along Parameterized Rendering Dimensions

For image synthesis, we used David Brainard's Vitrual World Toolbox, in combination with RenderToolbox4 and Mitsuba physically-based renderer (Heasly et al., (2014),`https://github.com/RenderToolbox/RenderToolbox4`, `http://www.mitsuba-renderer.org/`). We generated 4 sets of images, each based on 1 base image and several modified images. We modified the image along one rendering dimension until we achieved a unit length difference vector between the modified and base image (measured in luminance values) analogously to our eigen-distortion procedure. The modifications are loosely classified into two classes. The first class, object-centric distortions, includes rotation of the central object in each scene, as well as changing the brightness of the object itself, independent of it's surroundings. For the second class, image capture distortions, we translated the camera laterally, and zoomed the camera in towards the center of the scene. For each image set, we also computed the eigen-distortions for the On-Off model. Loosely, we classify the eigen-distortions generated by the model as adaptive structural changes.

To quantify human sensitivity to distortions along each of the above mentioned dimensions, we utilized the same psychophysical task described in Chapter 2. To do so, we used

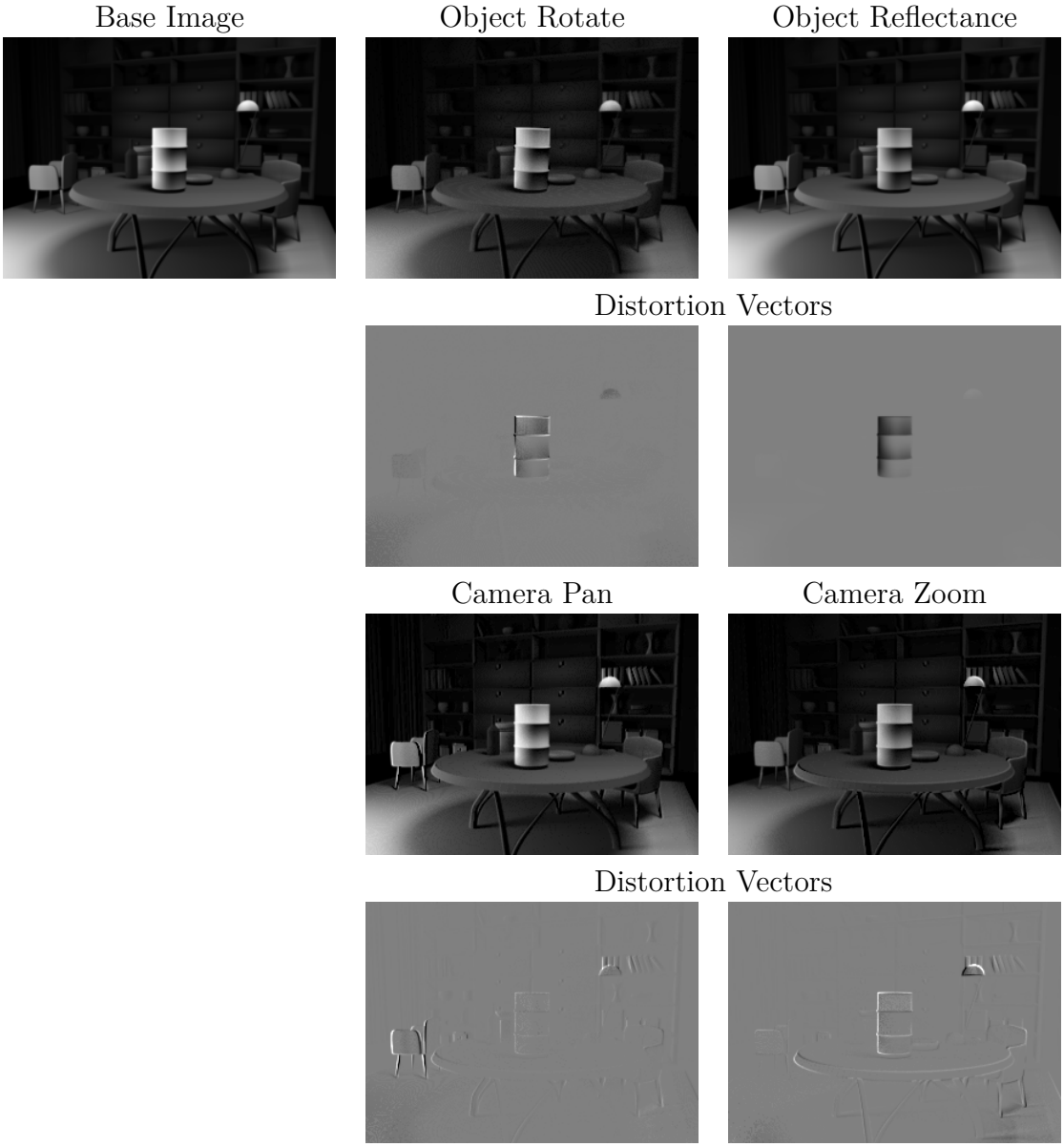

Figure 4.5: Distortions along the canonical dimensions described above. Images are best viewed in a display with luminance range from 5 to 300 cd/m^2 and a γ exponent of 2.4. **Top**: Object-centric distortions. Original image ($\vec{x}$), and sum of this image with each of the distortions. All distortion image intensities are scaled by the same amount ($\times 7$). **Second row**: Distortion Vectors for each modification. **Third and fourth rows**: Same, for the Image-Capture distortions. Distortion image intensities are scaled the same ($\times 7$).

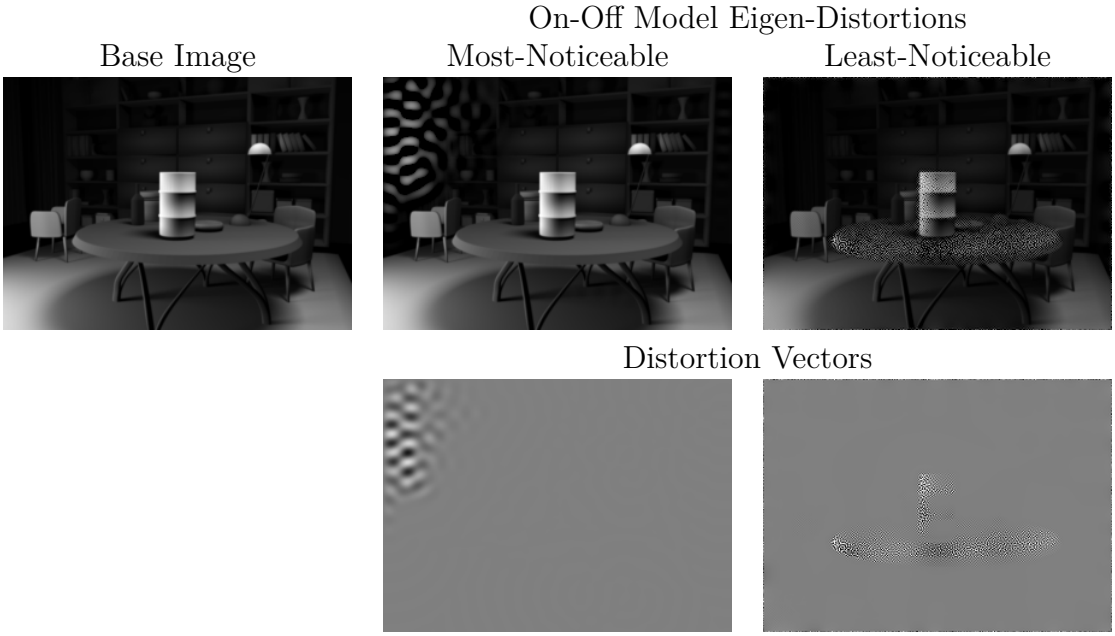

Figure 4.6: Eigen-distortions synthesized from the On-Off model. Images are best viewed in a display with luminance range from 5 to 300 cd/m^2 and a γ exponent of 2.4. **Top**: Eigen Distortions + Base Image. Original image ($\vec{x}$), and sum of this image with each of the distortions. Most-noticeable distortions scalle $\times 7$, least-noticeable distortion scaled $\times 20$. **Second row**: Distortion Vectors for each modification.

each unit length difference vector computed from the difference between the base image and one of the modified images as a distortion vector, $\vec{u}$, and scaled the vector amplitude α, until we located the human detection threshold. We estimated human thresholds for detecting these distortions using a two-alternative forced-choice task. On each trial, subjects were shown (for one second each with a half second blank screen between images, and in randomized order) a photographic image (15 degrees across), $\vec{x}$, and the same image distorted using one of the distortions, $\vec{x} + \alpha\hat{u}$, and then asked to indicate which image appeared more distorted. This procedure was repeated for 120 trials for each distortion vector, $\hat{u}$, over a range of α values, with ordering chosen by a standard psychophysical staircase procedure. The proportion of correct responses, as a function of α, was fit with a cumulative Gaussian function, and the subject's detection threshold, $T_s(\hat{u}; \vec{x})$ was estimated as the value of α for which the subject could distinguish the distorted image 75% of the time (See Appendix B for details). The images utilized in this section are $290x210$ pixels, and thus detection thresholds for this section (α parameters) are not directly comparable to chapter 2 (in which the images were $386x512$ pixels).

This linear approximation to distortions along each of these dimensions only holds for small distortion amplitudes, however, we find that the approximation holds across the regime in which we are testing. In fact, our method can be thought of as a Taylor expansion along our chosen distortion directions (object rotation, object reflectance change, lateral camera movement, camera zoom), similar to the method of Wang & Simoncelli, (2005). To measure human sensitivity to larger suprathreshold deviations along these dimensions, we suggest that the perceptual geodesic method of Hénaff, Goris and Simoncelli presents the most promising approach (Hénaff et al., (2017) and Hénaff et al., (2018)).

Model as observer

In addition to measuring human perceptual sensitivity to the above distortions, we also analyzed the log-likelihood that distances measured within the OnOff model produced the observed psychophysical data for each individual class of distortion. This analysis allows us to better understand what types of structural distortions the model is capturing well, and what types of structural distortion it weights differently than humans (see Appendix B for details). This analysis can point us towards the most promising directions to efficiently expand and improve our model.

4.1.2 Preliminary Results

In our preliminary experiments, we gathered data from 10 human subjects, across four sets of images. We begin by analyzing human detection thresholds for each of the rendering distortions averaged across subjects and images (See Figure 4.7). The most salient result is that the Adaptive structural distortions (the OnOff model's eigen-distortions), have the smallest and largest thresholds of the distortions that we measured. Of the class of rendering distortions, all are significantly less detectable than white noise of equivalent vector length. Contrary to what we expected, there is not a clean separation between object-centric distortions (rotation and object reflectance) and image capture distortions (Camera Zoom and Pan). This is somewhat surprising, as the distortions for the object-centric class are much more localized (like the most-noticeable Eigen-distortions of the OnOff model), but are still not more salient than the diffuse image capture class. This result, however, fits nicely into the framework we have developed. Despite the fact that these distortions are spatially localized, they represent natural, non-structural, transformations of the image, and so human observers are more tolerant to these modifications. We would like to note,

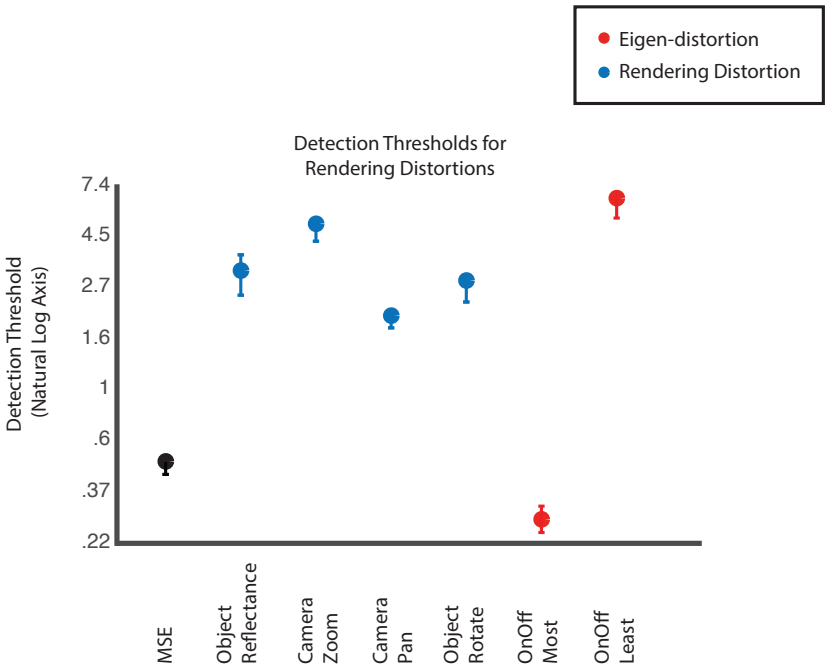

Figure 4.7: Average Detection Thresholds for Rendering Distortions show that the rendering distortions we tested are all less detectable than the OnOff model's most-noticeable distortion, and less-detectable than the model's least-noticeable distortion. Contrary to our expectations, there is not a clear separation within the rendering distortions between object-centric and image capture distortions.

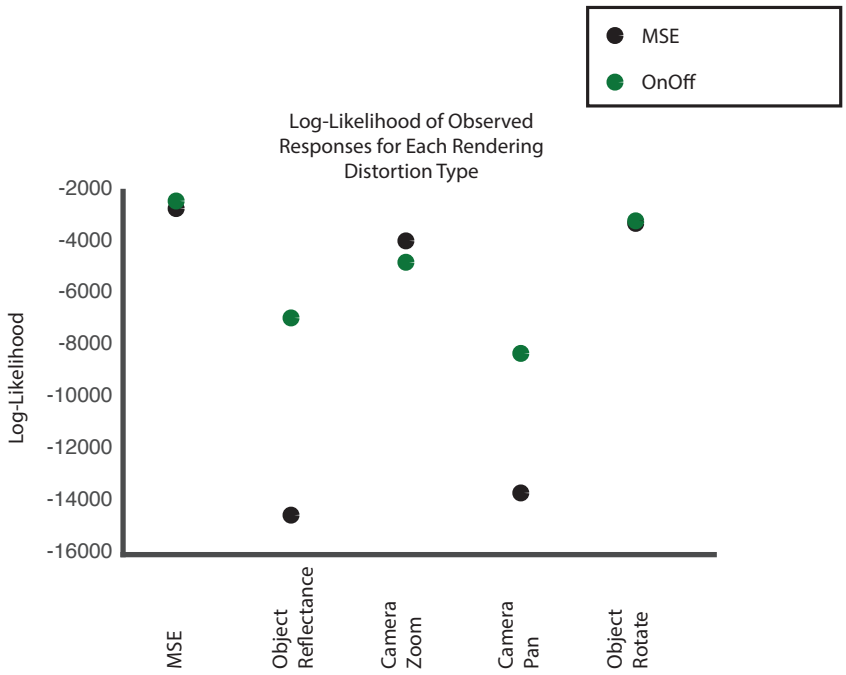

Figure 4.8: Log-Likelihood Analysis of OnOff Model Distances for Each Rendering Distortion (See Appendix B for details). While the OnOff model explains the responses to a subset of the rendering distortions better than MSE (Object Reflectance and Camera Pan), it performs equal to, or worse than, MSE at explaining human responses to the other three distortion classes. These results can mostly be explained by the model's oversensitivity to local translation. This result suggests that building a model of v1-complex cells on top of the OnOff representation would help the model capture local translation invariances that it currently is lacking.

however, that this is very preliminary data. These results are collected from a small batch of images, and may to not hold for every image.

In addition to measuring the average detection thresholds for each distortion class, we also analyzed the log-likelihood that the OnOff observer model produced the empirical psychophysical results for each of the rendering distortions and compared this to the log-likelihood that these results were produced by distances measured by pixel MSE (See Figure 4.8). This analysis shows that while OnOff distances explain some of these rendering distortion classes better than MSE (Object Reflectance and Camera Pan distortions), they underperform MSE at explaining Camera Zoom sensitivity, and are statistically indistinguishable from MSE for Object Rotation distortions. This result is somewhat unsurprising, the OnOff model is a model of very early vision, and it is unable to capture more complicated forms of image deformation. Both MSE and the OnOff model lack translation invariance of any kind, and will drastically over-predict human sensitivity to even small misalignments between elements of the original and distorted images. This results suggests that building local spatial invariance into the OnOff representation is an important direction for future work.

4.1.3 Future Work

Taken as a whole, the results in this thesis suggests that the nonlinear computations carried out in the early processing areas of the human visual system cannot be discounted or subsumed into the linear computations that begin with the first cortical area, V1. This computations create an adaptive Euclidean metric on the space of images, and are useful as a stand-in for human observers during complicated optimizations where fidelity to human perception is desired. Even so, these computations alone are not enough to fully capture

human perceptual sensitivity. There are many ways to improve upon our model. The most obvious of which is to expand it to be able to process color information. The next most obvious direction for further work, highlighted by results in this chapter, is to build in other known invariances of the human visual system, such as local translation invariance. Because our models are constructed as feedforward neural networks that loosely match the physiology of the first processing stages of the human visual system, the obvious way to build this invariance into the model is to cascade a model of V1 complex cells on top of the OnOff model outputs. When constructing and fitting this model, it will be important to carefully consider the nonlinearities at each stage. Another direction forward will be to test our models on more naturalistic tasks. Human vision is not constructed for viewing static images, but for natural sequences of static images. If the models are creating a metric space that appropriately captures human perception, it should be able to generalize to the sequential presentation of images. Though not explicitly constructed for this purpose, Hénaff and Simoncelli recently analyzed a cascaded neural model, constructed by stacking a model of V1 complex cells on top of our normalized model of the LGN. These authors showed that each stage of the mode significantly reduces the perceptual curvature of the trajectory of natural scenes in the same way as human subjects (private communication of unpublished data). They similarly showed that several classes of deep neural networks, including VGG, actually increase the curvature of natural trajectories at each stage, indicating a similar disconnect from Human perception for this class of models as shown in this thesis.

In addition to showing the importance of the nonlinear computations of early visual processing, we developed several toolsets that will be valuable to build upon in future work. In chapter 2, we explored the use of multidimensional Fisher information as a

toolset for exploring the high-dimensional sensitivity landscape of models. We derived these tools under several different model assumptions, related these tools to psychophysics, and showed how they may be used to build better model-driven psychophysical studies. These tools, however, are local approximations to the sensitivity of a model, meaning that their predictions do not hold over large distances. It will be exciting to relate these tools to the tools developed by Hénaff and Simoncelli for measuring and capturing supra-threshold, long-distance, perceptual deviations using perceptual geodesics. In addition, the noise assumptions we used in this thesis are fairly simple (a single stage of Gaussian or Poisson noise) (Hénaff et al., (2017) and Hénaff et al., (2018)). The reality of noise in actual neural circuits is more complicated. Neural responses can be described as modulated Poisson, or neurons with both independent Poisson variability and a multiplicative modulator shared across neurons (Goris et al., (2014)). This significantly complicates the computation of Fisher Information. Additionally, we have been approximating the function of stacked neural modules as deterministic, with a single stochastic module at the output of the final stage. In reality, neurons at every stage of the stack are noisy. This also creates difficulty for computing Fisher Information, as we now need to integrate over the internal noise as well as output noise. To properly capture how noise cascades through the system, we will have to develop methods to estimate the Fisher Information at the output stage using sampling methods.

In chapter 3, we developed a general purpose rendering framework for perceptually optimizing any rendered image under any display constraints. While this framework is useful as a benchmark, in practice, the optimization takes too much time to be utilized in real-world applications ($\sim$ 60 seconds per image). In future work, we would like to develop one-shot algorithms that approximate the results of this optimization under different

constraints with one forward pass through a model (or neural network). In addition, we would like to generalize the framework to optimize a sequence of images, concurrent with the development of metrics that capture sensitivity appropriately for a progression of natural images in sequence. We also want to test the performance of our metrics under other modern applications, and compare them to the performance of other metrics.

Finally, we would like to explore and understand the geometry of cascaded local divisive normalization. Several recent results suggest that a generalized form of divisive normalization can be utilized as a powerful tool for learning complicated image representations, and cascaded divisively normalized representations can be used to create state-of-the-art image compression algorithms (Ballé et al., (2017)) using many fewer layers that traditional cascaded LN- type neural networks. We believe this presents a potentially powerful tool for machine learning that is currently underutilized.

Appendix A : Model Implementation and Parameters

LGN models

We start by describing the simplest LGN model, the LN model. As our LGN models are a nested class, will build each successively more complicated model from the components already described for the simpler models.

LN model

Our LN model is constructed from a difference-of-gaussians filter with 31x31 support (one Gaussian representing the center receptive field, C, and one representing the surround receptive field, S), applied convolutional to the image, $\vec{x}$. Each of the gaussians is centered on the central pixel of the filter support, u, and is parameterized by a single parameter, σ controlling its variance, and constructed as the outer product of two equivalent 1-D gaussians. The gaussian representing the center receptive field is constructed as follows:

$$\vec{C} = \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{-\frac{(x-u)^2}{2\sigma_C^2}}$$

$$C = \vec{C}\vec{C}^T$$

The gaussian representing the surround receptive field is constructed as follows:

$$\vec{S} = \frac{1}{\sqrt{2\pi\sigma_S^2}} e^{-\frac{(x-u)^2}{2\sigma_S^2}}$$

$$S = \vec{S}\vec{S}^T$$

For all models except for the OnOff model, we constrain our center-surround filters to be of the On-center variety. Those filters are constructed as follows:

$$CS = C - .8S$$

The surround is weighted less than the center to match observed physiology and the valule .8 is fixed to reduce the number of learned parameters. Off-center filters are constructed by subtracting the center receptive field from the surround.

We convolve this filter with the input image, $\vec{x}$.

$$\vec{y} = CS \otimes \vec{x}$$

For all models, we rectify the output coefficients, $\vec{y}$, using a rectifying nonlinearity. Across all models, we utilize the "softplus" rectifier:

$$\vec{y}^{\,+} = log(1 + e^{\vec{y}})$$

Where $\vec{y}^{\,+}$ are the rectified output coefficients.

LG Model

The linear filtering stage of the LG model is constructed in the same way as the LN model filter. Prior to the rectifying nonlinearity, however, the LG model's linear filter coefficients, $\vec{y}_{Linear}$, go through a stage of divisive luminance normalization. A gaussian luminance pooling filter, L, is constructed in the same manner as the gaussian filters (C or S) above, parameterized by its variance, σ_L. We convolve the filter, L, with the image $\vec{x}$, to obtain a

luminance map, Lum.

$$Lum = L \otimes \vec{x}$$

And divide the output coefficients from the linear stage, $\vec{y}_{Linear}$, by this luminance map weighted by a learned parameter, α.

$$\vec{y}_{Lum} = \frac{\vec{y}_{Linear}}{1 + \alpha Lum}$$

As above, we rectify the outputs, $\vec{y}_{Lum}$ with a softplus nonlinearity to obtain the final outputs, $\vec{y}^{+}_{Lum}$.

LGG Model

The LGG model is constructed on top of the outputs of the LG model. We first convolve the squared output coefficients of the LG stage, $\vec{y}^{2}_{Lum}$ with a gaussian "contrast pool" filter, Con, (constructed in the same manner as above, and parameterized by a single variance parameter, σ_{Con}) in order to obtain a contrast map.

$$Contrast = \sqrt{Con \otimes \vec{y}^{\,2}_{lum}}$$

We divide the output coefficients of the LGG stage by this contrast map weighted by a learned parameter, β.

$$\vec{y}_{Con} = \frac{\vec{y}_{Lum}}{1 + \beta Contrast}$$

As above, we rectify the outputs, $\vec{y}_{Con}$, with a softplus nonlinearity to obtain the final outputs, $\vec{y}^{+}_{Con}$.

Models	σ_c	σ_s	σ_L	α	σ_{Con}	β
LN	.5339	6.148	NA	NA	NA	NA
LG	1.962	4.235	4.235	14.95	NA	NA
LGG	.7363	48.37	170.99	2.94	2.658	34.03
OnOff_{On}	1.237	30.12	76.4	3.26	7.49	7.34
OnOff_{Off}	0.3233	2.184	2.184	14.4	2.43	16.74

Table 2: **LGN Model Parameters**: Here we include the optimized parameters for each of the LGN models we tested in this thesis.

OnOff Model

The OnOff model is constructed as two parallel LGG models with independent parameterizations. The only enforced difference between the channels is that one of the linear filters, CS_{off}, is constructed to be an Off-center center-surround filter.

$$CS_{off} = S_{off} - .8C_{off}$$

Model Parameters

Below, we include the optimized parameters for each of the LGN models tested in this thesis (See Table 2). These can be used, along with the formulations above, to reconstruct the models.

Converting from SRGB to Linear with Luminance

We process all images in the luminance values they produce on the screen, in order to better approximate the signal that human subjects are actually seeing. In order to convert from SRGB pixel values that are linear with luminance, we utilize the following function. We first divide the SRGB pixels, $\vec{x}_{SRGB}$ by the maximum SRGB value, 255. We then convert

to linear-with-luminance values, $\vec{x}_L$.

For values in $\vec{x}_{SRGB}$ that are $\leq .04045$, we linearly remap these pixels by dividing their values by 12.92. For values in $\vec{x}_{SRGB}$ that are $> .04045$, we remap them as follows:

$$\vec{x}_L = \left(\frac{\vec{x}_{SRGB} + .055}{1.055}\right)^{2.4};$$

In order to properly display the images, we need to convert back to SRGB values. To do so, we undo the transformation above. For values in $\vec{x}_L$ that are $\leq .0031308$, we linearly remap these pixels by multiplying their values by 12.92. For values in $\vec{x}_L$ that are $> .0031308$, we remap them as follows:

$$\vec{x}_{SRGB} = \left(1.055 * \vec{x}_L\right)^{\frac{1}{2.4}} - .055$$

Appendix B : Estimating Human Perceptual Thresholds

To determine human perceptual thresholds for each eigendistortion, $\vec{e}_n$, we perform a staircased 2AFC (3 down, 1 up) in which we adjust variable α in the equation $\vec{x}_n + (\alpha\vec{e}_n\ )$. We model the human subject's performance on the 2AFC task as follows: The probability of a correct response, $P(c)$, is the product of the probability of a correct response given that the subject is paying attention to the task, $p(c|task)$, and 1 minus the probability that the subject will lapse, $p(lapse)$, plus the probability of a correct response despite lapsing, $p(c|lapse)$.

$$p(c) = p(c|task)(1 - p(lapse)) + p(c|lapse)p(lapse)$$

Where $p(c|task)$ is defined as a zero mean cumulative gaussian function defined by the amplitude parameter, α, and a variance parameter, σ, which defines the width of the gaussian, and which we will fit to data.

$$p(c|task) = cdf(\alpha, \mu, \sigma), \quad \mu = 0$$

$$cdf(\alpha, \mu, \sigma) = \frac{1}{2}\left[1 + erf(\frac{\alpha - \mu}{\sigma\sqrt{2}})\right]$$

Since there are only two choices at each response stage, the probability of a correct response despite lapsing is:

$$p(c|lapse) = \frac{1}{2}$$

And we fit the probability that the subject lapses on any given trial, γ, to the data.

$$p(lapse) = (\gamma)$$

Thus, the probability of a correct response given the parameters, α, γ and σ, is

$$p(c|\alpha, \gamma, \sigma) = [1 - \gamma]\ cdf(\alpha, 0, \sigma) + (\frac{\gamma}{2})$$

For each sample of size n, the probability of k observed correct responses during an entire trial is modeled as a the product of N indpendent Bernoulli processes:

$$p(k|\alpha, \gamma, \sigma) = \prod_{i=1}^{N} \binom{n_i}{k_i} p_i(c|\alpha, \gamma, \sigma)^{k_i} (1 - p_i(c|\alpha, \gamma, \sigma))^{(n_i - k_i)}$$

The log likelihood of the observed responses, k, given parameters, γ, σ and α is :

$$log(\mathcal{L}(k|\alpha, \gamma, \sigma)) = \sum_{i=1}^{N} \Big(log\binom{n_i}{k_i} + k_i log(p_i(c|\alpha, \gamma, \sigma)) + (ni - ki) log(1 - p_i(c|\alpha, \gamma, \sigma)) \Big)$$

We first fit a cumulative gaussian function to our data by maximizing the likelihood that the obeserved responses were produced given a single variance parameter, σ^*, and a single lapse parameter, γ^*, for each stimulus presented for a given subject.

$$(\gamma^*, \sigma^*) = \arg\max_{\gamma, \sigma} log(\mathcal{L}(k|\alpha, \gamma, \sigma))$$

We then estimate the threshold, $T(\vec{e}_n)$, as the amplitude, α, which gives a $p(c)$ of 75%, given the fit cdf:

$$T(\vec{e}_n) = cdf^{-1}(.75, 0, \sigma^*)$$

If our estimate of the lapse parameter $\gamma^* > .2$, indicating a high lapse rate, we exclude that set of trials from our dataset.

Estimating Log-Likelihood of Observed Psychophysical Data Given Model Distances

To determine the likelihood that distances measured within a model produced the observed psychophysical results, we first transform amplitude parameters, α, into model distances, D_m.

$$D_m(\vec{x}, \vec{e}, \alpha) = ||f_m(\vec{x}) - f_m(\vec{x} + \alpha\vec{e})||_2$$

We estimate the log likelihood that distances for each model can explain the observed correct responses under a single noise variance parameter.

$$p(c|task) = cdf(D_m, 0, \sigma_m)$$

We maximize the log likelihood over the variance parameter, σ_m, for the same formulation as above (with lapse parameter γ fixed at 0), and compare the log-likelihoods obtained from each model at the optimal parameters, σ_m^* across all stimuli presented across subjects. Larger likelihood indicates a higher correspondence between model derived distances and measured human response.

$$(\sigma_m^*) = \arg\max_{\sigma_m} log(\mathcal{L}_m(k|D_m, \sigma_m))$$

Estimating Log-Likelihoods for Each Individual Renderinyg Distortion Class

To determine the likelihood that distances measured within a model produced the observed psychophysical results for each distortion class, we first transform amplitude parameters, α, into model distances, D_m.

$$D_m(\vec{x}, \vec{e}, \alpha) = ||f_m(\vec{x}) - f_m(\vec{x} + \alpha\vec{e})||_2$$

We fix the noise variance parameter, σ, to the optimal noise variance parameter, σ^*_m estimated from the eigen-distortion data. This ensures that we are testing each model's ability to generalize to distortion types that have not been used to train any of its parameters.

We evaluate the log-likelihood of the observed responses for the distance data, D_m, given the previously fit optimal noise variance parameter σ^*_m.

$$log(\mathcal{L}(k|D_m, \sigma^*_m)) = \sum_{i=1}^{N} log\binom{n_i}{k_i} + k_i log(p_i(c|D_m, \sigma^*_m)) + (ni - ki) log(1 - p_i(c|D_m, \sigma^*_m))$$

Larger likelihood indicates higher correspondence between model derived distances and measured human response, indicating that the model captures this class of distortions well. Low likelihood indicates areas where the model may be improved.

Appendix C : Database Evaluation of Multi-scale IQA metrics

We compared the ability of each of our mutli-scale LGN based metrics to explain several widely utilized databases of image quality assessment with several state-of-the-art metrics. We report both Pearson and Spearman correlation, as is convention in the field. All correlations are computed over the gray-scale versions of images within each database (images transformed from RGB values to raw luminance values to match what the human viewer is actually seeing displayed on the screen). The authors wish to note that our analysis in preceding chapters suggests that these results on their own do not necessarily predict how a model generalizes outside of these databases. However, they do provide some information, and are the standard measure within the community.

Here, we refer to the version of the normalized laplacian pyramid distance (NLPD) with parameters estimated from image statistics as NLPD V1, and the version with parameters fit to the TID 2008 database as NLPD V2.

Table 3: Evaluation of IQA methods in different databases (Larson & Chandler, (2010), Ponomarenko et al., (2009), and Ponomarenko et al., (2015)). Pearson correlation and Spearman correlation (in parentheses) of distance metrics vs. human perceptual judgments. Numbers were obtained using the gray-scale version of the images in databases (see the text for details).

	TID 2008		TID 2013		CSIQ	
PSNR	0.52	(0.55)	0.64	(0.67)	0.76	(0.81)
SSIM	0.74	(0.78)	0.77	(0.80)	0.79	(0.87)
V1	0.81	(0.82)	0.81	(0.81)	0.87	(0.87)
MS-SSIM	0.79	(0.85)	0.79	(0.86)	0.77	(0.91)
VDP 2.2	0.80	(0.85)	0.78	(0.84)	0.90	(0.92)
NLPD V1	0.86	(0.87)	**0.88**	(**0.88**)	**0.92**	(0.92)
NLPD V2	**0.89**	(**0.89**)	**0.88**	(**0.88**)	0.90	(**0.93**)

Both of our multi-scale LGN metrics (NLPD V1 and NLPD V2) consistently outperform all of the tested state of the art metrics on these databases, in both absolute correlation, and rank-order correlation.

References

Ballé, J., Laparra, V. & Simoncelli, E. (2017). “End-to-end Optimized Image Compression”. In: *ICLR 2017*, 1–27.

Ballé, J., Laparra, V. & Simoncelli, E. P. (2016). “End-to-end optimization of nonlinear transform codes for perceptual quality”. In: *Proc. of 32nd Picture Coding Symposium*.

Barlow, H. (1961). “Possible principles underlying the transformation of sensory messages"”. In: *Sensory Communication*.

Bell, A. J. & Sejnowski, T. J. (1997). “The "independent components" of natural scenes are edge filters”. In: *Vision Research* 37(23), 3327–3338.

Burt, P. J. & Adelson, E. H. (1983). “The Laplacian Pyramid as a Compact Image Code”. In: *IEEE Trans. Communication* 31(4), 532–540.

Carandini, M. & Heeger, D. J. (2012). “Normalization as a canonical neural computation”. In: *Nature Reviews Neuroscience* 13.

Cerdá-Company, X., Párraga, C. A. & Otazu, X. (2016). “Which tone-mapping operator is the best? A comparative study of perceptual quality”. In: *arXiv e-prints* 1601.04450.

Chichilnisky, E. J. (2001). “A simple white noise analysis of neuronal light responses”. In: *Network: Computation in Neural Systems* 12, 199–213.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. (2016). “Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence”. In: *Scientific Report* 6.

Dodge, S. & Karam, L. (2017). “A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions”. In: *arxiv.org*.

Dosovitskiy, A. & Brox, T. (2016). “Generating Images with Perceptual Similarity Metrics based on Deep Networks”. In: *NIP2 2016: Neural Information Processing Systems*.

Eckert, M. P. & Bradley, A. P. (Nov. 1998). “Perceptual quality metrics applied to still image compression”. In: *Signal Processing* 70, 177–200.

Enroth-Cugell, C. & Pinto, L. (1970). “Algebraic Summation of Centre and Surround Inputs to Retinal Ganglion Cells of the Cat”. In: *Nature* 226, 458–459.

Eskicioglu, A. M. & Fisher, P. S. (Dec. 1995). “Image quality measures and their performance”. In: *IEEE Trans. Communications* 43, 2959–2965.

Fairchild, M. D. *The HDR Photographic Survey*. `http://www.cis.rit.edu/fairchild/HDR.html`.

Fattal, R. (2014). “Dehazing using Color-Lines”. In: *ACM Transaction on Graphics* 34(13) (1).

Ferwerda, J. A., Pattanaik, S. N., Shirley, P. & Greenberg, D. P. (1996). “A Model of Visual Adaptation for Realistic Image Synthesis”. In: *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

Fisher, R. (1925). “Theory of Statistical Estimation”. In: *Proceedings of the Cambridge Philosophical Society* 22, 700–725.

Floyd, R. W. & Steinberg, L. (1976). “An Adaptive Algorithm for Spatial Greyscale”. In: *Proc. Society for Information Display* 17(2), 75–77.

Freeman, J. & Simoncelli, E. P. (2011). “Metamers of the ventral stream”. In: *Nature Neuroscience* 14, 1195–1201.

Fukushima, K. (1980). “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. In: *Biological Cybernetics* 36, 193–202.

Girod, B. (1993). "What's wrong with mean-squared error". In: *Digital Images and Human Vision*, 207–220.

Goodfellow, I., Shlens, J. & Szegedy, C. (2014). "Explaining and Harnessing Adversarial Examples". In: *ICLR 2014*.

Goris, R., Movshon, J. & Simoncelli, E. P. (2014). "Partitioning Neural Variability". In: *Nature Neuroscience* 17(6), 858–865.

He, K., Sun, J. & Tang, X. (2011). "Single Image Haze Removal Using Dark Channel Prior". In: *IEEE Trans. Patterna Analysis and Machine Intelligence* 33(12), 2341–2353.

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B. & Brainard, D. H. (2014). "RenderToolbox3: MATLAB tools that facilitate physically based stimulus rendering for vision research". In: *Journal of Vision* 14(2).

Heeger, D. J. (1992). "Normalization of cell responses in cat striate cortex". In: *Visual Neuroscience* 9(2).

Hénaff, O., Goris, R. & Simoncelli, E. P. (2017). "Perceptual Straightening of Natural Video Trajectories". In: *Annual Meeting, Vision Sciences Society* 17.

Hénaff, O., Goris, R. & Simoncelli, E. (2018). "Perceptual Straightening of Natural Videos". In: *Computational and Systems Neuroscience (CoSyNe)*.

Hénaff, O. J. & Simoncelli, E. P. (2016). "Geodesics of learned representations". In: *ICLR 2016*.

Hodgkin, A. L. & Huxley, A. F. (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve". In: *Journal of Physiology* 117, 500–544.

Hoefflinger, B., ed. (2007). *High-Dynamic-Range (HDR) Vision*. Springer-Verlag.

Hornik, K. (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4(2), 251–257.

Hubel, D. H. & Wiesel, T. (1959). "Receptive fields of single neurones in the cat's striate cortex". In: *Journal of Physiology* 148, 574–591.

Hyvärinen, A. & Oja, E. (2000). "Independent Component Analysis: Algorithms and Applications". In: *Neural Networks* 13, 411–430.

Ioffe, S. & Szegedy, C. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ICLR 2015*.

Johnson, J., Alahi, A. & Li, F. F. (2016). "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *ECCV: The European Conference on Computer Vision*.

Karklin, Y. & Simoncelli, E. P. (2011). "Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons". In: *Advances in Neural Processing Systems (NIPS)* 24.

Khaligh-Razavi, S.-M. & Kriegeskorte, N. (2014). "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation". In: *PLOS Computational Biology* 10(11), e1003915.

Kingma, D. P. & Ba, J. L. (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv e-prints* 1412.6980.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *NIPS 2012: Neural Information Processing Systems*, 1–9.

Laparra, V., Berardino, A., Ballé, J. & Simoncelli, E. (2017). "Perceptually Optimized Image Rendering". In: *Journal of the Optical Society of America A* 34(9), 1511–1525.

Laparra, V., Muñoz-Marí, J. & Malo, J. (2010). "Divisive Normalization Image Quality Metric Revisited". In: *J. Optical Society of America A* 27(4), 852–864.

Laparra, V., Ballé, J., Berardino, A. & Simoncelli, E. P. (2016). "Perceptual image quality assessment using a normalized Laplacian pyramid". In: *Proc. SPIE, Human Vision and Electronic Imaging XXI*.

Larson, E. & Chandler, D. (2010). *Categorical Image Quality (CSIQ) Database*. `http://vision.okstate.edu/csiq`.

LeCun, Y. (1985). "Une procédure d'apprentissage pour Réseau à seuil assymétrique". In: *Cognitiva 85 : a la Frontière de l'Intelligence Artificielle, des Sciences de la Connaissance et des Neurosciences [in French]*, 599–604.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). "Deep Learning". In: *Nature* 521, 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. (1989). "Handwritten digit recognition with a back-propagation network". In: *Advances in Neural Processing Systems (NIPS)* 2, 396–404.

Lyu, S. & Simoncelli, E. P. (2008). "Nonlinear image representation using divisive normalization". In: *Proc. Computer Vision and Pattern Recognition*.

Ma., K., Liu, W., Liu, T., Wang, Z. & Tao, D. (Aug. 2017). "dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs". In: *IEEE Transactions on Image Processing* 26(8), 3951–3964.

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z. & Zuo, W. (2018). "End-to-End Blind Image Quality Assessment Using Deep Neural Networks". In: *IEEE Transactions on Image Processing* 27(3), 1202–1213.

Malo, J., Epifanio, I., Navarro, R. & Simoncelli, E. (2006). "Nonlinear image representation for efficient perceptual coding". In: *IEEE Transactions on Image Processing* 15.

Mannos, J. L. & Sakrison, D. J. (1974). "The effects of a visual fidelity criterion on the encoding of images". In: *IEEE Trans. Information Theory* 4, 525–536.

Mante, V., Frazor, R., Bonin, V., Geisler, W. & Carandini, M. (2005). "Independence of luminance and contrast in natural scenes and in the early visual system". In: *Nature neuroscience* 8(12), 1690–1697.

Mante, V., Bonin, V. & Carandini, M. (2008). "Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli." In: *Neuron* 58(4), 625–638.

Mantiuk Rafałand Daly, S. & Kerofsky, L. (2008). “Display Adaptive Tone Mapping”. In: *ACM Trans. Graph.* 27(3), 68:1–68:10.

Marmarelis, P. & Naka, K.-I. (1972). “White-Noise Analysis of a Neuron Chain: An Application of the Wiener Theory”. In: *Science* 175, 1276–1278.

Marmarelis, V. (1978). *Analysis of Physiological Systems The White-Noise Approach*. 1st ed. Springer.

Mises, R. von & Pollaczek-Geiringer, H. (1929). “Praktische Verfahren der Gleichungsauflösung”. In: *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik* 9, 152–164.

Mittal, A., Moorthy, A. K. & Bovik, A. C. (2012). “No-reference image quality assessment in the spatial domain”. In: *IEEE Trans. Image Process*, 4695–4708.

Mittal, A., Soundararajan, R. & Bovik, A. C. (2013). “Making a ’Completely Blind’ Image Quality Analyzer.” In: *IEEE Signal Process. Lett.* 20(3), 209–212.

Movshon, J., Thompson, I. & Tolhurst, D. (1978). “Spatial summation in the receptive fields of simple cells in the cat’s striate cortex”. In: *Journaly of Physiology* 283(53-77).

Narwaria, M., Mantiuk, R. K., Da Silva, M. P. & Le Callet, P. (2015). “HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images”. In: *J. Electronic Imaging* 24(1), 010501(1–3).

Nguyen A. Yosinski, J. & Clune, J. (2015). “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition”. In: *IEEE CVPR*.

Olmos, A. & Kingdom, F. A. A. (2004). “A biologically inspired algorithm for the recovery of shading and reflectance images”. In: *Perception* 33, 1463–1473.

Olshausen, B. A. & Field, D. J. (1996). “Emergence of simple-cell receptive field properties by learning a sparse code for natural images”. In: *Nature* 381, 607–609.

Pappas, T. N. & Safranek, R. J. (2000). “Perceptual criteria for image quality evaluation”. In: *Handbook of Image and Video Proc.*

Paris, S., Hasinoff, S. W. & Kautz, J. (2011). “Local Laplacian filters: edge-aware image processing with a Laplacian pyramid”. In: *ACM Trans. Graph.* 30(4), 68.

Parthasarathy, N., Batty, E., Falcon, W., Rutten, T., Rajpal, M., Chichilnisky, E. & Paninski, L. (2017). “Neural Networks for Efficient Bayesian Decoding of Natural Images from Retinal Neurons”. In: *Advances in Neural Processing Systems (NIPS)* 30.

Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D. & Greenberg, D. P. (1998). “A Multiscale Model of Adaptation and Spatial Vision for Realistic Image Display”. In: *Proc. 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH).*

Pattanaik, S. N., Tumblin, J., Yee, H. & Greenberg, D. P. (2000). “Time-dependent Visual Adaptation for Fast Realistic Image Display”. In: *Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH).*

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M. & Battisti, F. (2009). “TID2008 – A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics”. In: *Advances of Modern Radioelectronics* 10, 30–45.

Ponomarenko, N. N., Jin, L., Ieremeiev, O., Lukin, V. V., Egiazarian, K. O., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F. & Kuo, C.-C. J. (2015). “Image database TID2013: Peculiarities, results and perspectives”. In: *Sig. Proc.: Image Comm.* 30, 57–77.

Portilla, J. & Simoncelli, E. P. (2000). “A parametric texture model based on joint statistics of complex wavelet coefficients”. In: *Int’l Journal of Computer Vision* 40(1), "49–71".

Rumelhart, D. E., Hinton, G. & Williams, R. J. (1986). “Learning Internal Representations by Error Propogation”. In: *Nature* 323, 533–536.

Schwartz, O. & Simoncelli, E. P. (2001). “Natural signal statistics and sensory gain control”. In: *Nature Neuroscience* 4(8), 819–825.

Seriès, P., Stocker, A. A. & Simoncelli, E. P. (2009). “Is the Homunculus "Aware" of Sensory Adaptation?” In: *Neural Computation.*

Shapley, R., Enroth-Cugell, C., Bonds, A. B. & Kirby, A. (1972). "Gain Control in the Retina and Retinal Dynamics". In: *Nature* 236, 352–353.

Simoncelli, E. P. & Olshausen, B. A. (2001). "Natural Image Statistics and Neural Representation". In: *Annual Reivew Neuroscience* 24, 1193–1216.

Simoncelli, E. P., Paninski, L., Pillow, J. & Schwartz, O. (2004). "Characterization of Neural Responses with Stochastic Stimuli". In: *The New Cognitive Nuerosciences* 3.

Simonyan, K. & Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR 2015*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). "Intriguing properties of neural networks". In: *arXiv.org*.

Teo, P. C. & Heeger, D. J. (1994a). "Perceptual image distortion". In: *Proceedings of 1st International Conference on Image Processing*. Vol. 2, 982–986 vol.2.

Teo, P. & Heeger, D. J. (1994b). "Perceptual image distortion". In: *Proc. SPIE* 2179, 127–141.

Tumblin, J., Hodgins, J. K. & Guenter, B. K. (1999). "Two Methods for Display of High Contrast Images". In: *ACM Trans. Graph.* 18(1), 56–94.

Tumblin, J. & Rushmeier, H. (1993). "Tone Reproduction for Realistic Images". In: *IEEE Comput. Graph. Appl.* 13(6), 42–48.

VQEG (Mar. 2000). "Final report from the video quality experts group on the validation of objective models of video quality assessment." In: *http://www.vqeg.org/*.

Wang, Z. & Bovik, A. C. (Mar. 2002). "A universal image quality index". In: *IEEE Signal Processing Letters* 9, 81–84.

Wang, Z. (Dec. 2001). "Rate scalable foveated image and video communications". In: *PhD thesis, Dept. of ECE, The University of Texas at Austin*.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Trans. Image Processing* 13(4), 600–612.

Wang, Z. & Simoncelli, E. P. (2005). "An Adaptive Linear System Framework for Image Distortion Analysis". In: *Proceedings of the 12th IEEE International Conference on Image Processing* III, 1160–1163.

Wang, Z. & Simoncelli, E. P. (2008). "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual qualities". In: *Journal of Vision*.

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003). "Multi-Scale Structural Similarity for Image Quality Assessment". In: *37th Asilomar Conference on Signals, Systems and Computers*.

Watson, A. (Jan 2000). "Visual detection of spatial contrast patterns: Evaluation of five simple models," in: *Optics Express* 6, 12–33.

Watson, A. B. (1993). "DCTune: A technique for visual optimization of DCT quantization matrices for individual images". In: *Society for Information Display Digest of Technical Papers 24*, 946–949.

Watson, A. B. & Ahumada, A. J. (2005). "A standard model for foveal detection of spatial contrast". In: *Journal of Vision* 5(9).

Winkler, S. (1999). "A perceptual distortion metric for digital color video". In: *Proc. SPIE* 3644, 175–184.

Yamins, D. L. K., Hong, H., Cadieu, C., Solomon, E., Seibert, D. & DiCarlo, J. (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." In: *Proceedings of the National Academy of Sciences* 111(23), 8619–8624.

Yamins, D. L. K. & DiCarlo, J. J. (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19(3), 356–365.

Yamins, D., Hong, H., Cadieu, C. & DiCarlo, J. (2013). "Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream". In: *NIPS Neural Information Processing Systems*, 3093–3101.

Z. Wang, A. C. B. & Lu, L. (May 2002). "Why is image quality assessment so difficult". In: *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing* 4, 3313–3316.