Analysis and Optimization of Occluder-Based Imaging

by

Adam B. Yedidia

B.S., Massachusetts Institute of Technology (2014)
M.Eng., Massachusetts Institute of Technology (2015)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author . . . .
Department of Electrical Engineering and Computer Science
August 28, 2020

Certified by . . . .
Gregory W. Wornell
Sumitomo Professor of Engineering
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Analysis and Optimization of Occluder-Based Imaging

by

Adam B. Yedidia

Submitted to the Department of Electrical Engineering and Computer Science
on August 28, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Occluders, i.e. opaque objects, can be used in the apertures of cameras to supplement or replace a traditional lens. This thesis describes a novel mutual information-theoretic framework for analyzing and comparing occluders. It justifies the use of uniformly-redundant arrays (URAs), a popular choice of occluding pattern in coded-aperture imaging. This thesis shows these patterns to be optimal under ideal conditions using this framework. Outside of those ideal conditions, this thesis proposes a method for selecting between different URAs and compares it to other occluder-selection methods, such as a greedy search, identifying under which conditions each is preferable. It also shows, analytically and empirically, the superiority of designed occluding patterns like URAs to random occluding patterns. The mutual-information theoretic framework is compared to a similar, MSE-minimizing framework.

This thesis also considers the use of occluders in the context of non-line-of-sight (NLoS) imaging, used as "accidental cameras." The idea of the accidental camera is to opportunistically make use of occluding objects that happen to be available as ad-hoc coded apertures. Methods of this class, having originally been developed by Torralba and Freeman in 2012, are extended in this thesis to a wide variety of different scenarios, and used to solve formerly unsolved NLoS problems. These include imaging around a corner using the corner as the occluder, imaging a light-field of an unknown scene using a known, calibrated occluder, and imaging an unknown scene using an unknown occluder. The tools of the aforementioned framework are used to draw tentative conclusions about NLoS imaging systems, including resolution limitations due to longer light wavelengths and the quality of reconstructions across different systems.

Thesis Supervisor: Gregory W. Wornell
Title: Sumitomo Professor of Engineering

Acknowledgments

First and foremost, I'd like to thank Gregory Wornell, my research advisor, for his invaluable help, ideas, and for the home he gave me at MIT. Without him, I never would have been introduced to the research topic of non-line-of-sight imaging, which plays a central role in my thesis. Everybody always says that choosing an advisor is the most important part of graduate school; it warms my heart to say that they were absolutely right.

I'd also like to thank the rest of my thesis committee, Bill Freeman and Frédo Durand, who met with me regularly throughout my PhD and whose ideas, support, and discussions were tremendously useful. I'd like to especially thank Bill Freeman for his mentorship and generosity with his time and attention; it was thanks to him that I found a postdoctoral assignment during the dreadful time of COVID-19. Bill Freeman's boundless creativity will forever inspire me.

I'd like to thank Martin Rinard, who co-advised me at the start of my PhD, for his tremendous generosity, kindness, and liveliness. Martin gave me a home when I first came to graduate school, and he funded me before I was covered by my NDSEG fellowship. My research went in a different direction over the course of my PhD, but I have not forgotten his generosity, nor will I.

I'd also like to thank Phillip Stanley-Marbell, for taking me under his wing at the start of my PhD. He was ludicrously nice to me. I didn't know people could be that nice. He and Max Shulaker mentored me during my ill-fated foray into hardware design, and I will always remember their incredible kindness.

I'd like to thank Regina Barzilay for introducing me into her group, and for the opportunities she gave me to work with all the people at MGH.

I'd like to thank all my co-authors over the course of my PhD: Miika Aittala, Ganesh Ajjanagadde, Manisha Bahl, Manel Baradad, Regina Barzilay, Katie Bouman, Richard Brent, Frédo Durand, Bill Freeman, Connie Lehman, Andrea Lincoln, Nick Locascio, Lukas Murmann, Prafull Sharma, Christos Thrampoulidis, Antonio Torralba, Greg Wornell, Vickie Ye, and Lili Yu. A few of these I would like to thank in particular.

Richard Brent answered an email from me, a stranger emailing him out of the blue. Working with him was a great pleasure. I am extremely grateful to him.

Christos Thrampoulidis mentored me throughout the middle of my PhD. His advice and mentorship made me grow tremendously as a researcher—and even more importantly, Christos is a lot of fun to be around.

Ganesh Ajjanagadde and I had countless stimulating conversations. Ganesh is tremendously generous with his time and his ideas, and I could not have asked for a better person with whom to share my time in graduate school.

And then, of course, Andrea Lincoln. I'll get back to her.

I'd like to thank Tricia O'Donnell, for the endless things she did for me over the last five years. She was always there, always helpful, and always cheerful, even after I messed up.

I'd like to thank Janet Fischer, for all the help she's given me over the course of my PhD.

Most of my research was funded through a DARPA grant and an NDSEG fellowship. I am grateful to both organizations for making my research possible.

I am grateful to all my friends, from Epsilon Theta and from elsewhere, sadly too many to name. But Alex Arkhipov and David Farhi deserve particular mention for providing lots of useful advice about my research, some of which made it into this thesis. And Sam Bankman-Fried, as well, for giving me good advice about grad school and life in general.

I am deeply, deeply grateful to my family: Ann Blair, Jonathan Yedidia, and Zachary Yedidia, and the uncountable things they've done for me. I know how lucky I am.

Finally, I'd like to thank my fiancée, Andrea Lincoln. The story of my PhD—indeed, the story of most of my adult life—is in large part a story about her as well as me. She is the best part of my life, an uncomplicated good. She is part of who I am now, and I am deeply proud of that. I can't possibly describe all of what she is to me here—nor should I—but Andrea is my partner, lover, best friend, and confidante.

Relatedly, I'd like to thank Virginia Vassilevska-Williams and Ryan Williams. I know what they did for me, even if they may not.

Contents

List of Figures

Chapter 1

Introduction

What is a camera? The dictionary tells us that a camera is an “optical instrument to record images.” When you think of a camera, you probably think of something with a lens, like a digital camera or, more recently, a smartphone—or, in the future, perhaps something else entirely. Lenses are very effective imaging tools—indeed, even our eyes evolved to use them!—but cameras can be built without them. The first known camera that could create permanent photographs was developed by Joseph Nicéphore Niépce, who used a camera obscura, or pinhole camera, to project an image of a scene onto dark pewter coated with bright bitumen [44]. The bitumen, exposed to light, hardened and became insoluble, so that it would not wash off the pewter when rinsed with water, leaving the parts exposed to the light to stay white while the rest turned dark. See Figure 1-1 to see the earliest surviving photograph.

This photograph was created by a pinhole camera, which is part of a class of cameras that I call “occluder-based” cameras. An *occluder* is any object that blocks light; in the case of a pinhole camera, this object blocks nearly all of the light, except for that which is allowed to pass through a tiny hole. Pinhole cameras are special, because the image projected onto the pewter is identical to the scene, except that it’s spatially reversed and blurred in proportion to the size of the hole. That makes the design of a pinhole camera the *simplest* of the occluder-based cameras, because the projected image is easily interpreted, without any need for further computation. *Simplest*, however, does not mean *best*, and more recently-developed occluder-based

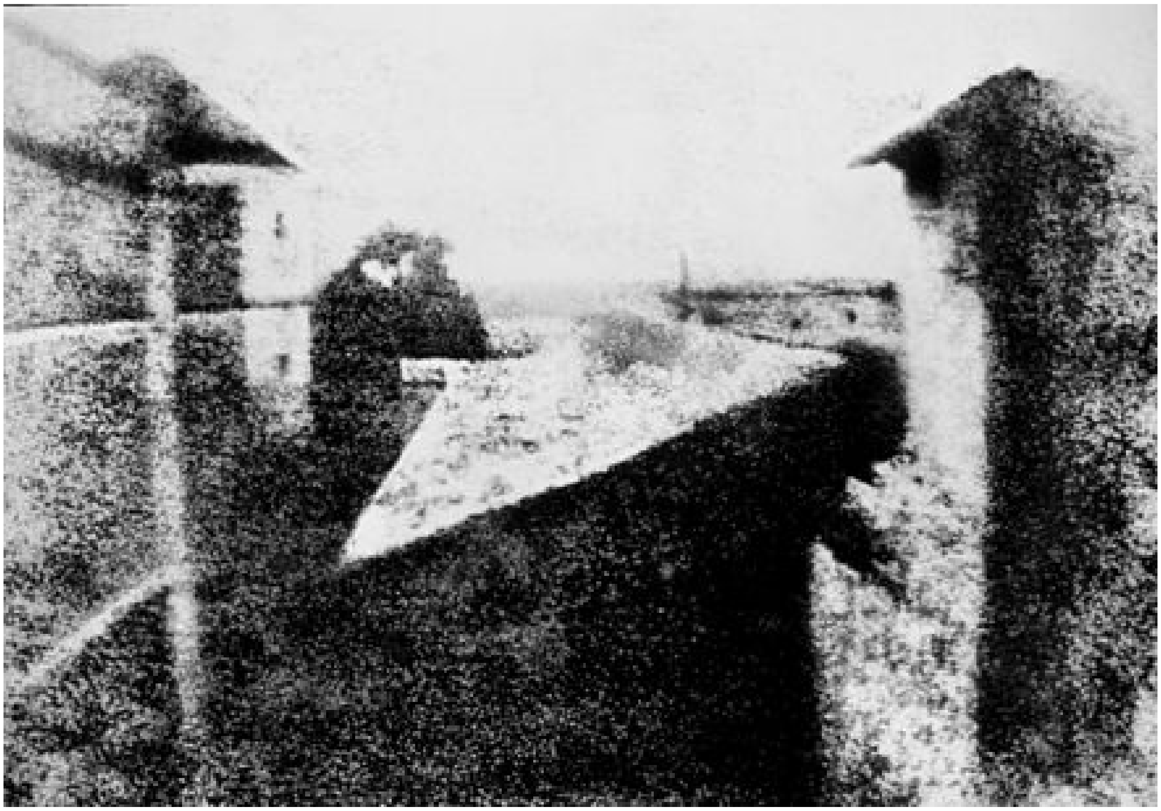


Figure 1-1: The earliest surviving photograph, taken by Joseph Nicéphore Niépce in 1825, using an occluder-based camera.

cameras [9, 30, 102, 41] have abandoned the pinhole in favor of more complicated patterns of occlusion. These more complicated patterns vary from implementation to implementation, though the most popular choice is a set of patterns called “uniformly redundant arrays.” Each of these non-pinhole patterns yield unintelligible projected images, but enable much more accurate computational reconstructions in the presence of noise. See Figure 1-2 for a side-by-side simulated comparison of the two methods.

One of the contributions of this thesis is provide a framework, based on information theory, that explains why uniformly redundant arrays are better than pinholes. Using this framework, I will describe the circumstances under which we know uniformly redundant arrays are the best option, and how to choose which uniformly redundant array to choose. This is what Chapter 3 is about: choosing the best occluder-based camera system possible, given a particular set of imaging circumstances. Occluder-based designer cameras aren’t the only topic of this thesis, however. In addition to designer cameras, this thesis contributes to the field of occluder-based *accidental cameras*.

Before I explain the concept of accidental cameras, let’s revisit the question we began with. What is a camera? A camera is an optical instrument to record images. What are cameras made of? Generally, cameras are the combination of two objects. The first is what I will eventually call the “observation plane,” or more simply the “observation.” This object is generally something photosensitive, like film, a digital light sensor, or pewter coated with bitumen as in our earlier example. The second is what I will eventually call the “aperture frame.” The purpose of this latter object is simply to make the observation non-uniform. It’s necessary to have something in between the scene and the observation, be it a lens or an occluder, because without it, our observation would be uniformly illuminated. This, in turn, means the observation would contain almost no information about the scene.

Accidental occluder-based cameras take a creative approach to both the observation and the aperture frame. In most of the accidental-camera systems described in this thesis, the observation plane is a blank wall or floor, recorded by a digital camera positioned elsewhere (the “observer”). The aperture frame is some occluding

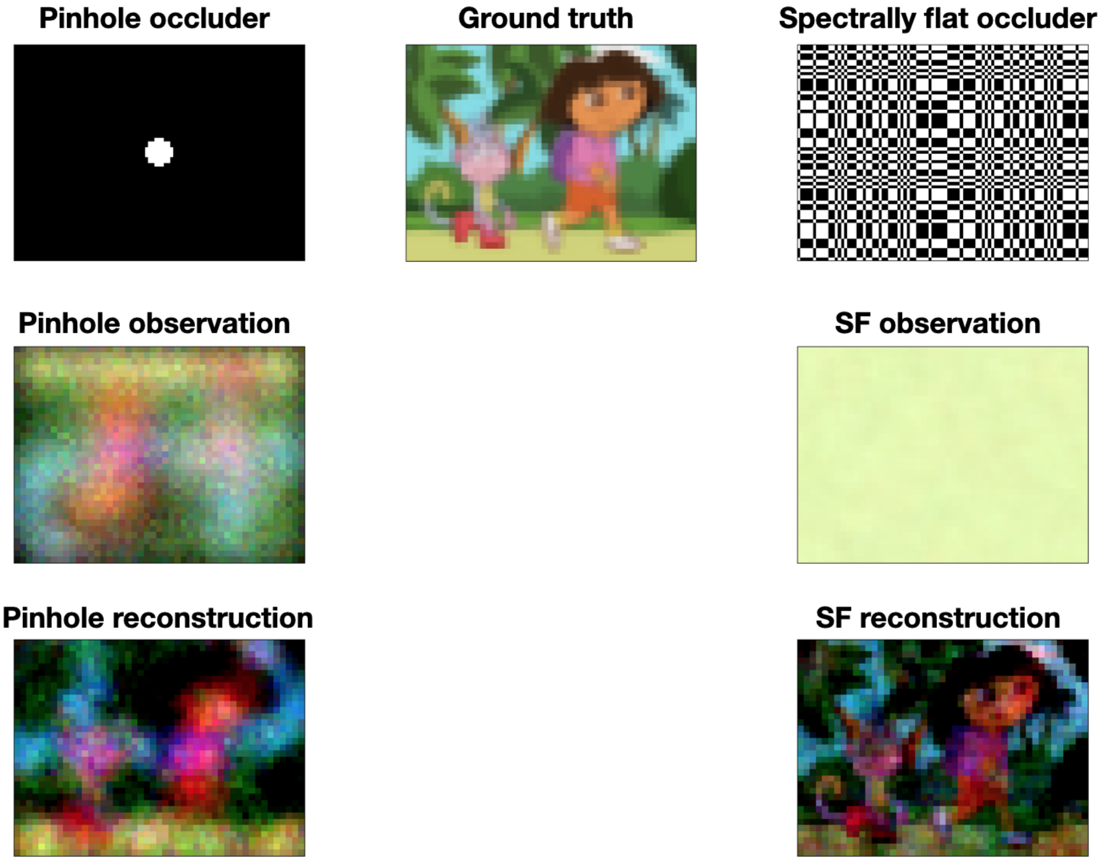


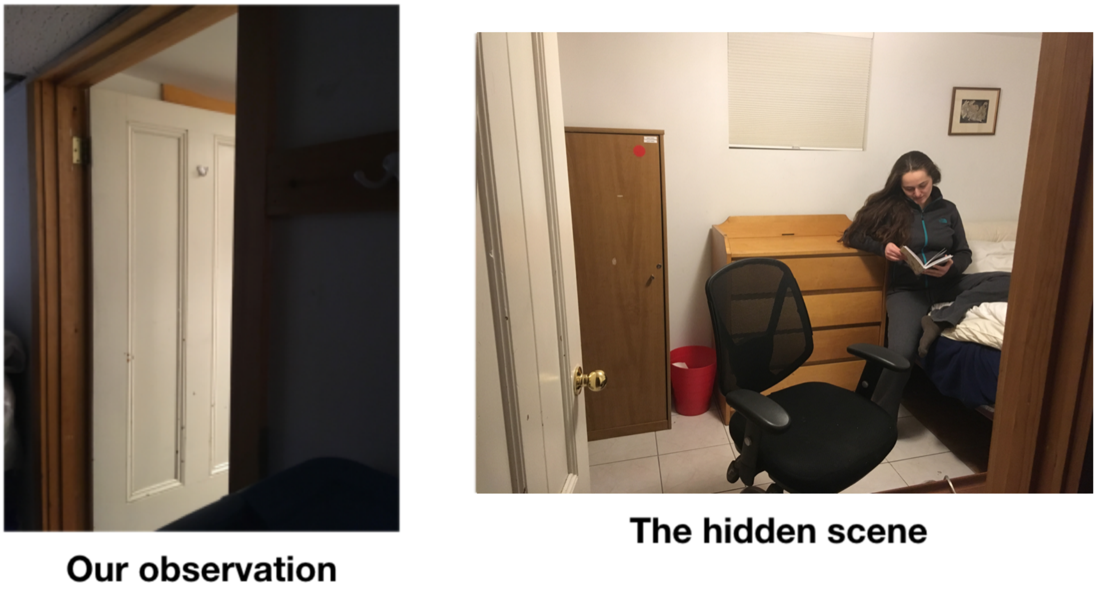
Figure 1-2: This figure will appear again, later in this thesis, when more context with which to understand it has been given. The point here is that when the simulated unknown scene (top middle) is projected through a pinhole occluder (top left), it results in a more intelligible projected image, or "observation" (middle left) but a lower-quality reconstruction of the scene (lower left). In constrast, when the simulated unknown scene is projected through a more complicated occluder, in this case a uniformly redundant array (top right), it results in a less intelligible projected image (middle right) but a higher-quality reconstruction of the scene (lower right). It makes sense that historically, the simpler (left) approach would be the first, because no computation is required to understand the observation, but the more sophisticated (right) approach is more powerful.

object, like a wall (see Section 4.1) or any opaque object (see Section 4.3 or 4.5). One possible purpose of an accidental occluder-based camera is non-line-of-sight (NLoS) imaging: in other words, to “see” things are not directly in view of the observer. To give a very basic example, imagine you are sitting in your room, with sunlight shining in from a window. From the square of sunlight on the floor, and the location of your window, you can infer in what direction the Sun is. In a certain sense, you have “seen” the Sun; without actually looking out your window, you know where it appears in the sky. Even if you don’t know exactly what you would see if you looked out your window, you know where the Sun would be in that picture.

The same method that you used to infer the position of the Sun without looking out your window, could in principle be used to infer the position of everything else visible from your window, without looking out your window. This is because it’s not just the Sun that emits light; visible objects are visible because they emit (reflected) light, it’s just that they are emitting much *less* light. But even if it’s too subtle to be seen with the naked eye, there *is* light streaming through your window that comes not directly from the Sun, but reflected off something outside, and if you were sharp-eyed enough to see it, you would be able to infer exactly what was outside your window, without actually looking out your window. Indeed, this is more or less exactly what was shown by Torralba and Freeman in 2012 [97], in one of the first publications to discuss occluder-based accidental cameras.

What good is it to use sophisticated computational techniques to infer what you *would* see if you *were* to look out a window? Why not just look out the window, like a normal person? There are scenarios in which being able to solve analogous problems can be very useful. For example, this style of imaging technique could let you “see” into a room you didn’t want to enter (see Figure 1-3). Perhaps even more usefully, occluder-based accidental cameras could be used to detect oncoming traffic around a corner, and help prevent life-threatening collisions (see Figure 1-4). Additionally, occluder-based systems can be used to create smaller cameras than lenses allow [9] or to perform X-ray imaging, which is often incompatible with lenses [1].

Chapters 1 and 2 provide background, and Chapters 3 and 4 provide contributions.



Our observation

The hidden scene

Figure 1-3: A hypothetical scenario in which occluder-based techniques could be used for non-line-of-sight imaging. On the left, we can see a door, but we don't know anything about what's in the room it leads to (shown on the right). Together, the door and the chair inside the room form an accidental camera. Using the method presented in Section 4.5, an onlooker could, in principle, try to reconstruct an image of the room by observing the door by using the occlusion provided by the chair and the motion provided by the person. This is true even though neither the chair nor the person is visible to the onlooker! I say "in principle" because in practice, the signal strength in the room as shown would be too weak to get a sharp reconstruction—this is meant as an illustration of what is possible with accidental cameras, not a description of current capabilities.

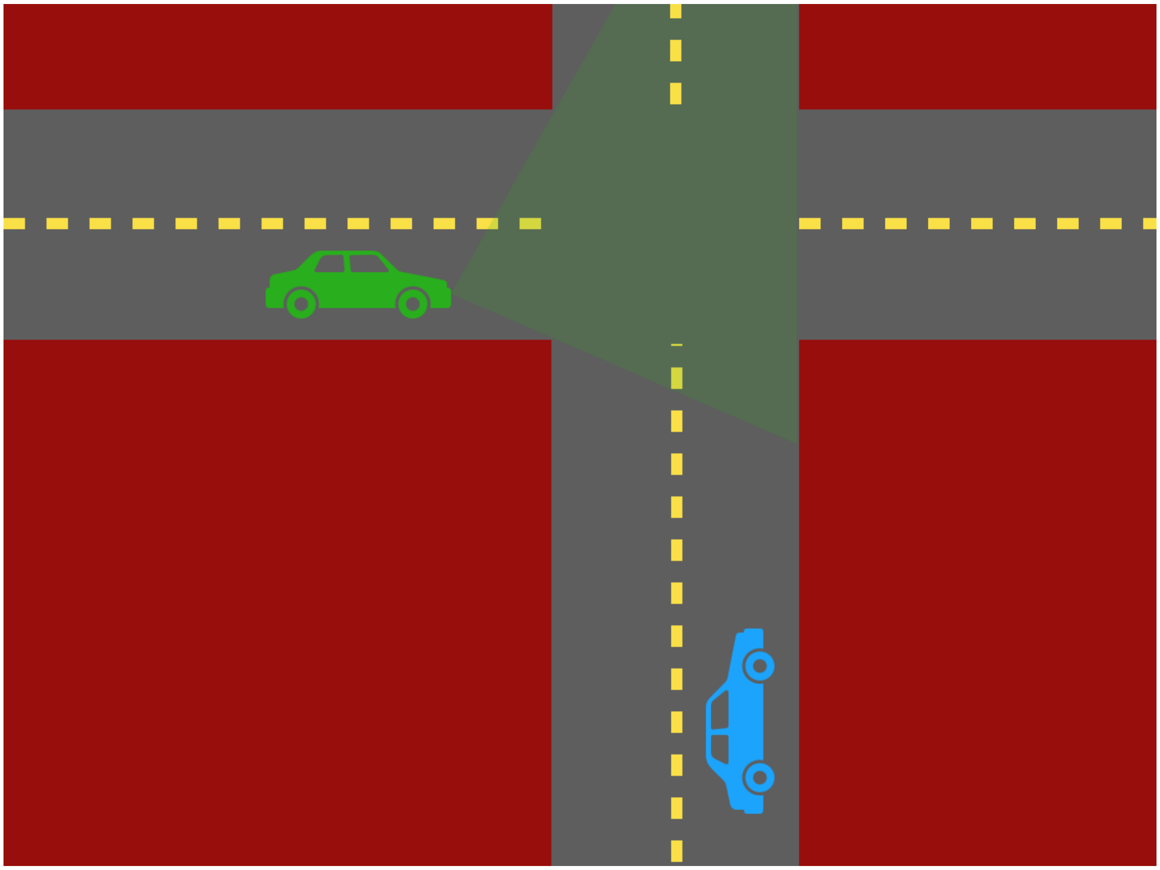


Figure 1-4: Imagine you are in the blue car, approaching the intersection ahead. There's a green car headed towards the intersection around the corner, but you can't see it, because there's a building blocking your view. However, the green car is reflecting green light onto the street in front of you, and that green light is something that *is* within your field of view, even if it's too faint to be seen with the naked eye. The intersection in front of you, combined with the building to your left, form an accidental camera. An automatic driving system could use it to infer the angle from the corner and the speed of the green car around the corner using the computational imaging techniques described in Section 4.1. Systems like these could be used to avoid a deadly collision, if the green car is about to run a red light.

The contents of Chapter 3 (which is broadly about finding the best possible occluder-based aperture frame) and those of Chapter 4 (which is about making use of whichever occluders happen to be handy in NLoS settings) are largely separate. However, the analysis of Chapter 3 is of obvious relevance to NLoS occluder-based imaging even when finding an optimal occluder "in the wild" isn't realistic.

The framework described and used in this thesis is a powerful tool for analyzing occluder-based imaging. Occluder-based non-line-of-sight imaging remains a promising area for future research, with the deep image prior being particularly promising, as discussed in Section 4.6.

1.1 Overview

This thesis is meant to serve a few purposes simultaneously. The first will be to provide a clear and easy introduction to the core concepts of occluder-based imaging; my hope is that anybody who wishes to learn more about the topic, either because they are interested in an application in non-line-of-sight imaging purposes, coded-aperture imaging, or some other application, will find this thesis to be helpful. I am also hopeful that this thesis can be understood even by readers with relatively little background; I intend for many of its sections to be comprehensible without any background in physics or optics, and only an undergraduate-level understanding of linear algebra, signal processing, and calculus (in decreasing order of importance).

Chapter 3 gives a complete treatment of which mask-based occluders are optimal, under a variety of different scenarios, as well as a novel framework for analyzing mask-based occluders. When optimality can't be proven, Chapter 3 is still able to give specific, helpful advice for which mask-based occluders are likely to perform best. The contents of Chapter 3 will be helpful for designing an occluding-mask-based camera. For other types of nontraditional cameras—cameras that use a different sort of aperture frame, perhaps, but still not a lens (one example is the DiffuserCam [8])—extending the analysis of Chapter 3 should prove straightforward.

As for Chapter 4, I recommend reading it to any reader whose primary research focus is in non-line-of-sight imaging, especially passive non-line-of-sight imaging. Chapter 4 describes how to solve non-line-of-sight imaging problems under a variety of constraints.

1.2 Ray Optics and BRDFs

Throughout my thesis, unless I say otherwise, I'll be using the *ray optics model* (also known as the geometrical optics model) of light. This means that, as a convenient abstraction, I'll be assuming that light moves in a straight line through the air, can be bent when it hits a different light-propagating medium (like a lens), and may be absorbed or reflected by the materials it hits (like a wall). Moreover, I'll be generally assuming that light intensity is additive, meaning that two rays of light hitting the same point will generate an intensity equal to the intensity that would be generated by the sum of each individual ray. This corresponds to assuming that the light we care about is incoherent and as such won't interfere with itself—a reasonable assumption when the light in question is coming from the sun or from a commercial electric light. Finally, using the ray optics model means that I'll be ignoring the effects of diffraction, which is reasonable when modeling the behavior of visible-wavelength incoherent light hitting macroscopic structures.

What happens when the light hits an opaque surface, according to this model? Some of it will be absorbed, and some reflected. How much of it is reflected, and in what directions, is described by the *bidirectional reflectance distribution function*, or *BRDF*, of the surface. In flatland (meaning a 2D world), the BRDF is a function of two variables, the angle of the incoming light ω_i and the angle of the outgoing light ω_r, and returns the ratio of the reflected radiance (the outgoing power per unit angle per unit projected area on the surface) along ω_r to the irradiance (the incoming power per unit area) incident on the surface from ω_i.

There are a variety of BRDFs that describe real-world surfaces, including the specular BRDF (which describe mirrors and other shiny surfaces well) and Phong BRDFs (due to [80], which describe glossy surfaces well). In this thesis, however, the focus is on the Lambertian BRDF, which is defined below:

$$f_{\text{Lambertian}}(\theta_{\text{in}}, \theta_{\text{out}}) = \rho/\pi$$

Surfaces that have this BRDF are often called *Lambertian*, *matte*, or *diffuse* surfaces, and I will use these terms interchangeably in this thesis. Intuitively, this BRDF takes the incoming light and emits it "equally in all directions." Once again ρ is a constant that determines the overall surface brightess. Note that the Lambertian BRDF is completely independent of the angle of the incident light—it emits the light it reflects in exactly the same way no matter where the light came from.[1]

1.3 The Paraxial Approximation

In this section, I introduce the paraxial approximation, a commonly used assumption in computational imaging [109, 69, 94, 87, 25]. The idea is that when a light source is far from a surface, its illumination of that surface will be approximately uniform.

1.3.1 A point light source and a nearby surface

Let's suppose we live in a 2D world of Lambertian surfaces and diffuse light sources. (When I say a "diffuse" light source, I mean that the light source emits light equally in all directions.) Consider a point light source suspended at $(0, y_p)$, with a Lambertian surface at $y = 0$ (see Figure 1-5). What pattern of illumination can we expect to see on the surface?

The way we proceed with this analysis is to discretize the surface into many small patches, and then to consider what fraction of the light radiating out of the point light source hits any single given small patch of the surface. We assume each patch is small enough that its apparent intensity is constant across the patch. Asking what fraction of light radiating out of the point light source hitting any given patch is

[1] For a 2D surface living in a 3D world, the Lambertian surface BRDF is exactly the same, but four arguments are required to fully parametrize the incoming and outgoing angles.

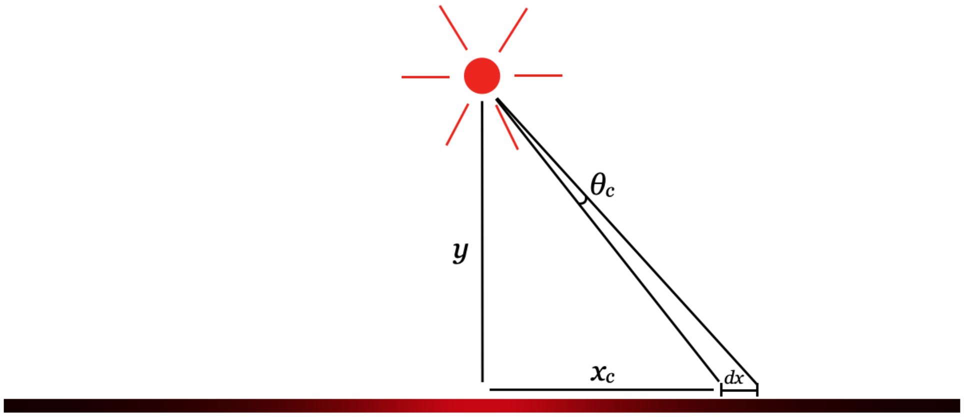


Figure 1-5: A diagram illustrating the following configuration: a point light source at $(0, y_p)$, with a Lambertian surface at $y = 0$. We are interested in the resulting illumination pattern on the Lambertian surface; to investigate it, we measure the illumination of a small patch on the surface that extends from $(x_c, 0)$ to $(x_c + dx, 0)$. The illumination of that patch will be proportional to the angle θ_c of the point source's light subtended by the patch.

equivalent to asking what angle over the light source is subtended by that patch, and then dividing that angle by 2π.

Supposing that the patch extends from $(x_c, 0)$ to $(x_c + dx, 0)$, trigonometry tells us that θ_c, the angle subtended by the patch, is given by:

$$\theta_c = \tan^{-1}\left(\frac{x_c + dx}{y_p}\right) - \tan^{-1}\left(\frac{x_c}{y_p}\right)$$

What happens as we consider increasingly smaller and smaller patches dx? The definition of the derivative tells us that $\lim_{dx \to 0} \theta_c = dx \cdot \frac{d}{dx_c}(\tan^{-1}(\frac{x_c+dx}{y_p})) = dx(y_p/(x^2 + y_p^2))$. Thus the luminance of a patch on the surface, assuming that the point source had a luminance of 1, would be $dx(y_p/(2\pi(x^2 + y_p^2)))$.We can say that the continuous illumination function of the surface $I(x)$ is the following:

$$I(x) = \frac{y_p}{2\pi(x^2 + y_p^2)}$$

This simple formula captures a lot of interesting phenomena. Consider for instance that we take $x = 0$, meaning we consider the illumination only of the closest point

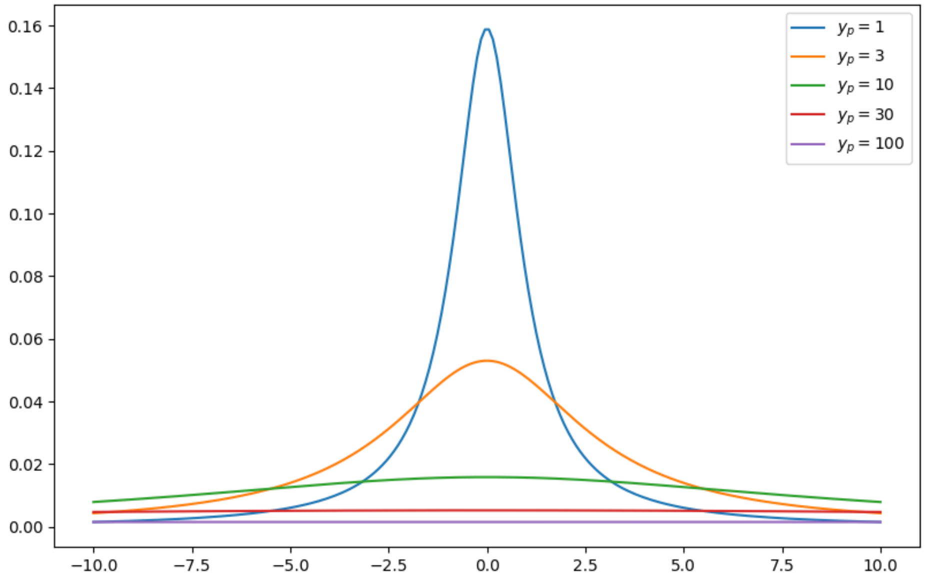


Figure 1-6: Different illumination patterns depending on different possible values of y_p, with the Lambertian surface extending from $x = -10$ to $x = 10$. As this plot shows, the paraxial approximation starts becomes reasonable around $y_p = 30$.

on the surface to the point source. The formula tells us then that the illumination of that point goes as $1/y_p$, meaning that it scales inversely with that point's distance from the point source. Now consider fixing $y_p = 1$ and varying x. This gives us an illumination pattern that scales with $1/(1+x^2)$. The closer the surface is to the point source (meaning a smaller y), the narrower the hump will be (See Fig. 1-6). Also note that no matter what y_p is, we have:

$$\int_{-\infty}^{\infty} \frac{y_p}{2\pi(x^2 + y_p^2)} dx = 1/2$$

It stands to reason that this is true, because no matter how far the surface is from the point source, if the surface is infinitely broad, exactly half the light from the point source will hit the surface. Additionally, for reference, I'll provide here the illumination function for the equivalent situation in three dimensions: a point source of luminance 1 suspended at $(0, 0, z_p)$, and a plane at $z = 0$. Then, the illumination function $I(x, y)$ can be derived in much the same way as in the two-dimensional case. This function is given by:

$$I(x, y) = \frac{z_p}{4\pi(x^2 + y^2 + z_p^2)^{3/2}} \tag{1.1}$$

The important thing at this point is that, as shown in Figure 1-6, the illumination pattern becomes flatter and broader the further the point source is from the surface. This phemonenon is what we rely on when we use the paraxial approximation. This is the assumption that the contribution of a point light source to a faraway surface is approximately constant across that surface. This assumption holds as long as the size of the surface in question is much smaller than the distance of the point source to the surface; that is, if, for all relevant values of x, $x^2 \ll y_p^2$, then it follows that $I(x)$ holds a constant value of approximately $1/(2\pi y_p)$ $(1/(4\pi z_p^2)$ in three dimensions), assuming the point source has a luminance of 1.

Because of the quadratic dependence on x and y_p in Eq. 1.1, the paraxial approximation is reasonable even when the difference between x and y_p isn't enormous; for example, if you hold a diffuse light source three meters away from the center of a flat surface two meters in diameter, the brightness of that surface won't vary by more than about 16% (compare $1/9^{3/2}$ to $1/10^{3/2}$). The paraxial approximation gets relied on very heavily, both in my research and in work by others, and admittedly the reason for that isn't that it's always a hugely robust assumption to real-world situations (after all, depending on the application, sometimes 16% can matter a lot!). The reason, rather, is that it's an extremely *convenient* assumption. For the time being I'll leave it at that, but in later sections we will see that tolerating the paraxial approximation grants us quite a lot of mathematical convenience.

1.4 The Standard Configuration

In this section, I will briefly describe what I call "the standard configuration," and introduce some terminology that I will use throughout the dissertation. The simplest version of the standard configuration is shown in Fig. 1-7: three parallel frames in

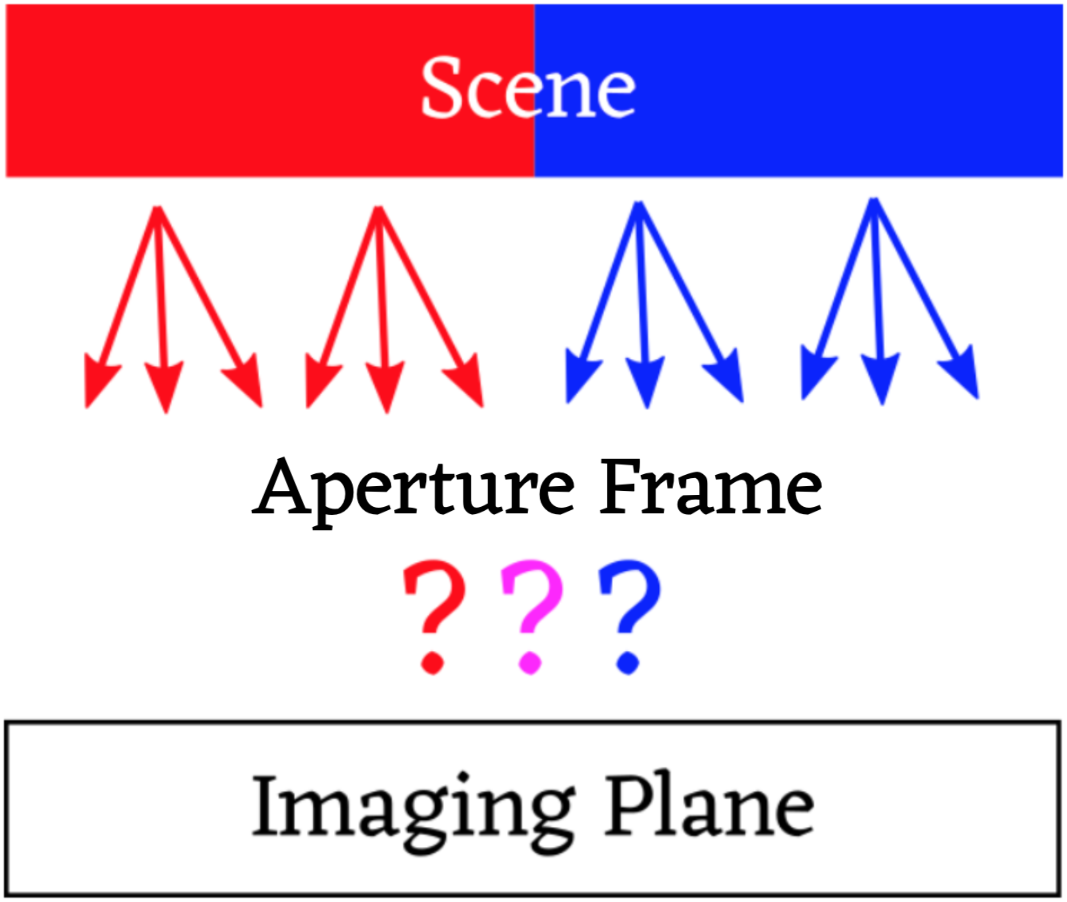


Figure 1-7: The standard configuration: three parallel frames in flatland, with the "aperture frame" halfway in between the scene and the observation plane. The aperture frame could be anything from an occluder, to a lens, to a random scattering pattern.

flatland, with the "aperture frame" halfway in between the scene and the observation plane. The presumption is that the observation is a known quantity, and we'd like to infer what's in the scene. Depending on the details of the problem, the aperture frame may also be a known quantity, or its form may be unknown. In any case, we'd like to see how much we're able to infer about the scene from the observation thanks to (or despite!) the presence of the aperture frame.

The term "aperture frame" is left deliberately vague. In an ordinary camera, the aperture frame would be a lens. In most of this dissertation, I'll be considering aperture frames that don't directly focus the light from the scene like a lens would, but partially occlude the scene. In principle, there are any number of other realistic aperture frames.

Of course, there are many other ways to relax the standard configuration to make it richer or more realistic. The aperture frame need not be halfway in between the observation plane; the three frames need not be parallel to each other; the scene need not be planar. And, of course, the real world isn't flatland (a term which I will use to refer to a 2D, rather than a 3D, world)! But the standard configuration is a great starting point for any optical analysis.

There are two more things about the way I model the standard configuration that must be mentioned here. The first is although the scene and observation plane will, in reality, be continuously varying objects, I will by default be assuming them to be n-dimensional vectors, with each entry being the intensity of one little piece of the observation. Obviously, this is an approximation of reality, since it assumes that the intensity is uniform across each little piece. As n gets larger, this approximation will become closer and closer to reality.

The second thing is that the standard way I index the scene vector is left-to-right, but the standard way I index the observation vector is **right-to-left**. (In a 3D world, I would reverse the labeling along both dimensions.) In principle, you could do everything I do in this thesis with the labeling of the observation going left-to-right, but it would make the analysis much less pleasant. See Figure 1-8 to see a side-by-side comparison of the true configuration and the modeled configuration.

1.4.1 The Transfer Matrix

Another critical concept in my dissertation is the *transfer matrix.* This is a standard concept in computational imaging, used when linearity can be established [103, 71, 72]. The transfer matrix is a matrix that describes the action of an aperture frame on the scene to create the observation. To be more precise, suppose we approximate the scene by a vector $\vec{x}$, where each entry of that vector gives the illumination of a single patch of the scene. Suppose that we approximate the observation plane in the same way with a vector $\vec{y}$. Then, the transfer matrix, A, will be whichever matrix satisfies

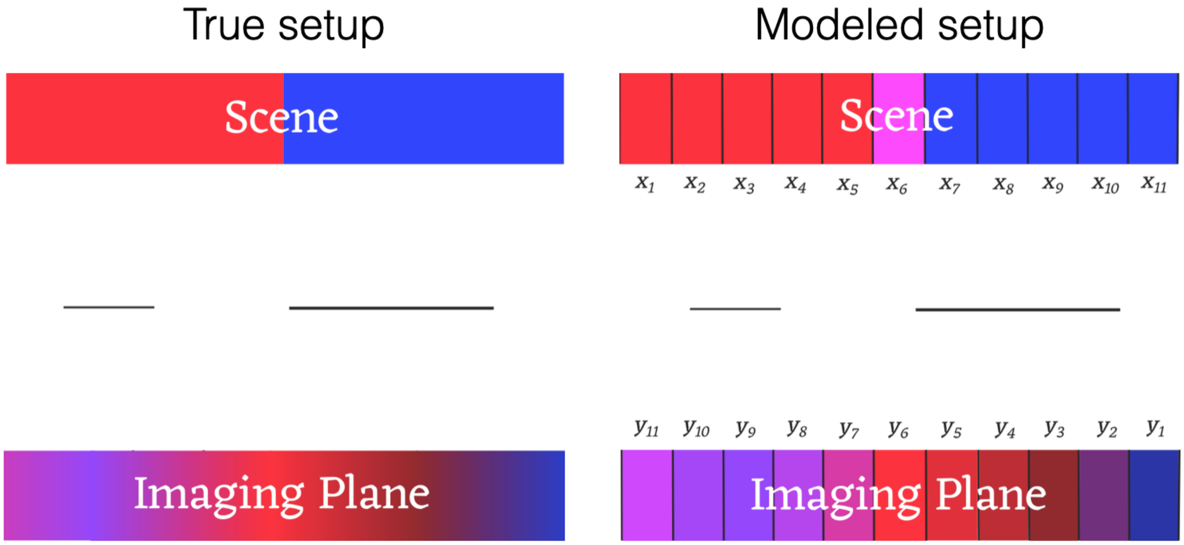

True setup
Modeled setup
Scene
Scene
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11}
y_{11} y_{10} y_9 y_8 y_7 y_6 y_5 y_4 y_3 y_2 y_1
Imaging Plane
Imaging Plane

Figure 1-8: Left: the true configuration. Right: the modeled configuration, with the resolution level being $n = 11$. Note the labeling is reversed on the observation relative to the scene. The faithfulness of the modeled configuration to the true configuration will get better and better as n increases.

$\vec{y} = A\vec{x}$ for all possible pairs $(\vec{x}, \vec{y})$.

How do we know that such a matrix even exists for all aperture frames? Well, if we accept the assumptions implicit in the ray-optics model described in Sec. 1.2—that is, we ignore the effects of diffraction and assume that light is incoherent—then what you observe should be a linear function of the presence or absence of light sources. That means that we call what you see if light a is on $f(a)$, and what you see if light b is on $f(b)$, then what you see when both lights are on, $f(a+b)$, should be the sum of what you saw in either case, $f(a) + f(b)$.

Because we're assuming that combining light sources behaves linearly, most real-world objects we can put in between a scene and an observation plane should be representable by a transfer matrix A. How do we actually construct this transfer matrix? Each column of the transfer matrix corresponds to the illumination pattern on the observation plane in response to an impulse light source at each different patch in the scene; see Figure 1-9. One can actually empirically measure the transfer matrix in real-world imaging systems using this method; Section 4.3, which describes

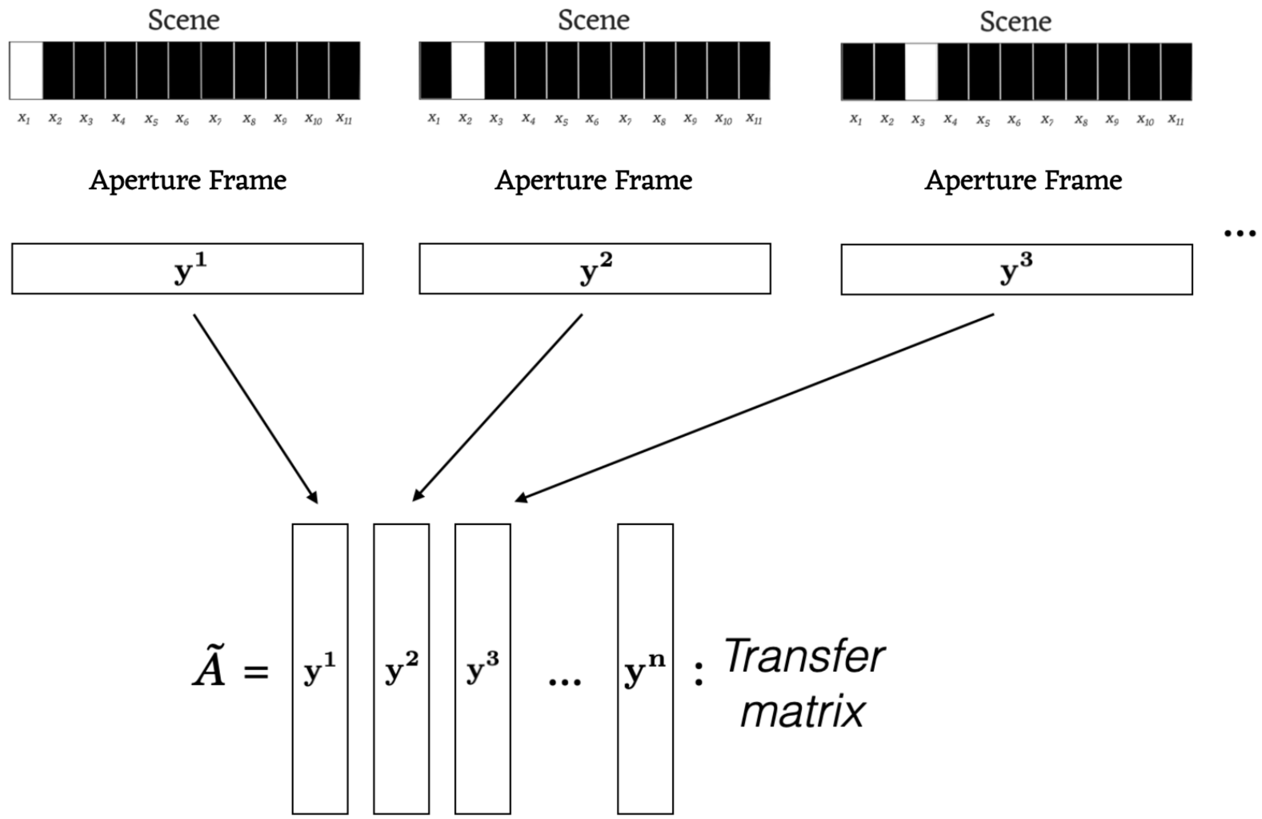

Scene
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11}
Aperture Frame
$\mathbf{y^1}$
$\mathbf{y^2}$
$\mathbf{y^3}$
...
$\tilde{A}$ = $\mathbf{y^1}$ $\mathbf{y^2}$ $\mathbf{y^3}$ ... $\mathbf{y^n}$: Transfer matrix

Figure 1-9: How the transfer matrix comes out of the imaging system: each column of the transfer matrix corresponds to the illumination pattern on the observation plane in response to an impulse light source at each different point in the scene.

a calibrated occluder-based imaging system, does exactly this!

Transfer matrices take a variety of different forms depending on what the aperture frame is and what it's doing; Figure 1-10 shows the action of a variety of example aperture frames, with an example transfer matrix corresponding to one of them. And from the example transfer matrix shown in Figure 1-10, we can now understand the choice to label the observation plane from right-to-left instead of left-to-right. The *bottom-most* rows of the transfer matrix shown correspond to what the *left-most* elements of the observation see, and correspondingly the top-most rows of the transfer matrix shown correspond to what the right-most elements of the observation see. In the transfer matrix we see in Fig. 1-10, the diagonals of constancy[2] go from upper-left to lower-right, but if we'd chosen the opposite labeling scheme, they would go

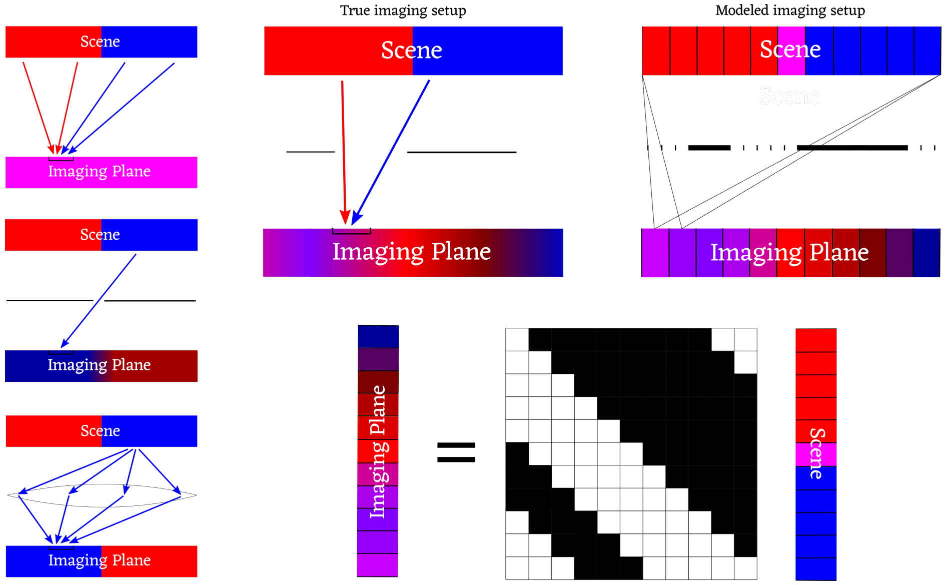


Figure 1-10: Three imaging systems (left column, top-to-bottom): no aperture, a pinhole and a lens. Arrows indicate paths light from the scene takes to a particular point on the imaging plane. On the right is an arbitrary mask, an illustration of its discretization and the corresponding transfer matrix.

from *lower*-left to *upper*-right. Transfer matrices with diagonals of constancy that go from upper-left to lower-right are called *Toeplitz* matrices, and they are well-studied and have many nice properties which will be helpful for the analysis that follows. A special class of Toeplitz matrices, called *circulant* matrices, which will be described later, have especially nice properties.

So suppose we have a scene x, an observation y, and a transfer matrix A that represents how the aperture frame distorts the scene to produce the observation. If $y = Ax$, and we know A and y, what transfer matrices A are best? In the absence of noise, any full-rank transfer matrix A should allow us in principle to perfectly reconstruct $x = A^{-1}y$. That makes the question of which transfer matrix not very interesting—it's a multi-way tie between all full-rank transfer matrices, which make up the vast majority of possible transfer matrices. This tie is broken in the presence

[2] The "diagonals of constancy" are the lines you can draw on certain matrices such that all values on that line have the same value. For a discrete matrix, they can only go in four directions: up to down, left to right, upper-left to lower-right, and upper-right to lower-left. However, for a continuous matrix, they could in principle go in any direction. This phenomenon gets discussed more in Section 3.11.

of noise.

1.4.2 The probability distribution over scenes

Of course, it's unrealistic to expect no noise. Every real-world imaging setting will have at least some noise, and in any case it's the presence of noise that makes the problem interesting, and helps distinguish between better and worse transfer matrices, even if both matrices have the same rank. The scene models described in this section closely match those of [109], though the explanation here is more detailed.

Adding noise, our new equation becomes:

$$y = Ax + \eta$$

where η is another vector representing random noise. Now that we have introduced a random variable into our equation, we will need to provide a probability distribution not only for η (to describe how the noise is distributed) but also for x.

Let's start by discussing the probability distribution of the scene vector, x. The simplest model to begin with is to have each entry of the scene be independent and identically distributed (IID), and drawn from a Gaussian. For example, suppose that each entry of the scene vector x was independently drawn from a Gaussian with a mean of μ and a standard deviation of σ.[3] To describe this situation, we can write:

$$x \sim \mathcal{N}(\mu I, \sigma^2)$$

Before we can proceed with this model, there are a few problems to worry about. The first is possible negative entries. Real scenes don't cast negative light! To solve this problem, we take $\mu \gg \sigma$. That way, the probability of negative entries will be vanishingly small. A vanishingly small chance of a negative entry is good enough; it

[3] This is, in fact, a realistic model in most passive imaging settings. In truth, the amount of light from a light source is a Poisson distribution, with the intensity of the light source as its mean; but in the limit of large mean, a Poisson distribution is well-approximated by a normal distribution. Any ambiently-lit room will have more than enough light to be at the point where normal distribution is an excellent approximation of the amount of light being emitted by the room's light sources.

means that our model's distance from the real world due to this issue (where negative entries are impossible) is also vanishingly small.

However, recall that x is a discrete vector, but it is meant to represent a continuous scene of fixed size. We haven't yet talked about the number of entries in x, which we'll call n. The variable n controls how finely we discretize the scene x. Ideally, choosing n to be larger will mean that our discrete representation of the true continuous scene will be more faithful (though perhaps at a computational cost). And we might also hope that once n gets large enough, that's a close enough approximation to the real scene that increasing it further won't make the model noticeably better. That's not such an unrealistic expectation; after all, if you're reading these words on a laptop screen, you're probably looking at a discrete array of a couple thousand by a thousand pixels, and that's plenty enough to give you the impression of a "continuous" image on your screen. Tripling the number of pixels on your laptop without increasing the size of your screen probably won't improve your impression of how "continuous" your screen looks by much, unless you're very good at noticing this kind of thing. We can convert a continuous scene to a discrete one simply by making each discrete patch of the scene contain the average intensity within that patch, with the discrete representation become increasingly faithful as the patches get smaller and smaller. (See Figure 1-8)

There's a problem with our model of x as stated, the probability distribution over scenes. The problem is this: varying n should give us representations of the true, continuous scene that are varyingly faithful. However, choosing a different value of n shouldn't qualitatively change what the scene looks like. It shouldn't change the underlying model of the continuous scene, only our representation of it.

So first, we need to make sure that the total luminance of the scene (in other words, the total amount of light the scene emits) doesn't depend on n. But we said earlier that each entry of x was IID with a mean of μ, so at the moment the expected total luminance of the scene is $n\mu$. This means that μ is going to have to depend on n; in particular, we'll say that $\mu = J/n$, where J is a constant that represents the total luminance of the scene.

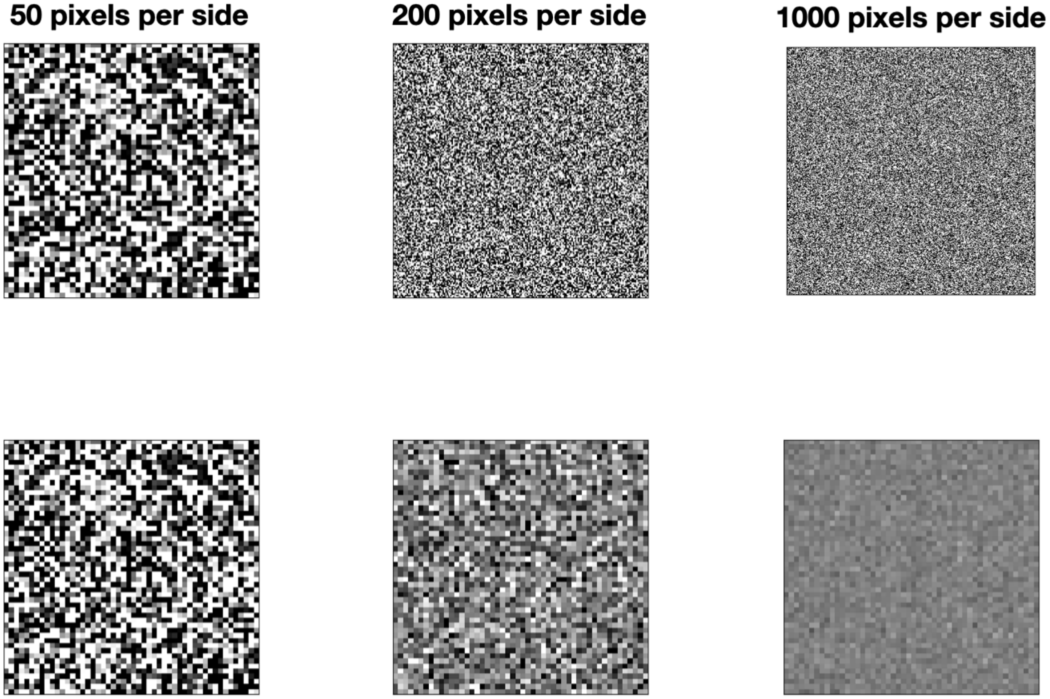


Figure 1-11: Top row: IID scenes with $\sigma = 1$, with three different levels of scene discretization ($n = 50$, $n = 200$, and $n = 1000$). Bottom row: each of those scenes, pixellated so that they have 50 pixels to a side. As you can see, the pixellated scenes look different from each other—this is a problem, because how finely we choose to discretize the scene shouldn't make a difference to what the scene looks like. This is why IID scenes of different levels of discretization are qualitatively different from each other, unlike correlated scenes whose covariance matrices are chosen carefully to scale properly with discretization level.

Additionally, there is the problem that of each entry of x is IID and drawn from a Gaussian, then instances of x for different values of n will be discrete representations of different underlying scenes. See Figure 1-11.

The solution to this issue is for our model of the scene to include correlations between nearby pixels. We can do this by supposing that the covariance matrix of the scene includes off-diagonal elements:

$$x \sim \mathcal{N}(\mathbf{Q}, \sigma^2)$$

The covariance matrix $\mathbf{Q}$ captures the correlations between nearby pixels. In real scenes, the closer together two pixels are, the more correlated they'll become. Our model should be faithful to this as well—and in doing so, we will simultaneously create the situation we wanted before, in which scenes with different values of n look qualitatively similar to each other, just at different levels of fidelity.

To make things concrete, I'll provide an example of a scene covariance matrix, which I'll call the *exponential-decay prior.* Recall that an IID covariance matrix would just be a multiple of the identity, $\mathbf{Q} = \frac{\theta}{n} I$, where θ is a constant in n. The exponential-decay prior is given by $\mathbf{Q} = \mathbf{F}_n^* \mathbf{D}^\star \mathbf{F}_n$, where $\mathbf{F}_n$ is the normalized DFT matrix of size n and $\mathbf{D}^\star$ is a diagonal matrix with the following entries: $d_1 = 1$, $d_i^\star = d_{n-i+1}^\star = \frac{\theta}{n} \beta^{\frac{i-1}{\lceil (n-1)/2 \rceil}}, i = 2, \ldots, \lceil (n+1)/2 \rceil$, for some frequency decay rate parameter $0 < \beta < 1$. A lower β implies a more strongly correlated scene. Scenes discretized to different levels (varying n), generated with an exponential-decay prior covariance matrix, will look qualitatively the same no matter what n is—a desirable property, because it means that no matter what we choose n to be, we will be talking about the different models for the same underlying continuous scene.

See Figure 1-12 for what these covariance matrices look like, and what scenes generated from them look like.

Note that even if the notion of an IID scene is incoherent as we take $n \to \infty$, an IID scene for a fixed value of n can still represent something realistic. It's just that if one wants to then represent that same probability distribution over scenes at a finer level of discretization, it will be necessary to introduce spatial correlations. Put another way, two IID scene models with different values of n can't represent the same underlying continuous scene, but they can each individually represent a model of a continuous scene that makes sense.

1.4.3 Noise

Now that we have a model of the probability distribution over scenes, we need a model for the probability distribution over the noise. This crucially depends on what application it is we care about. We'll try and make the noise model general enough to apply well to all cases.

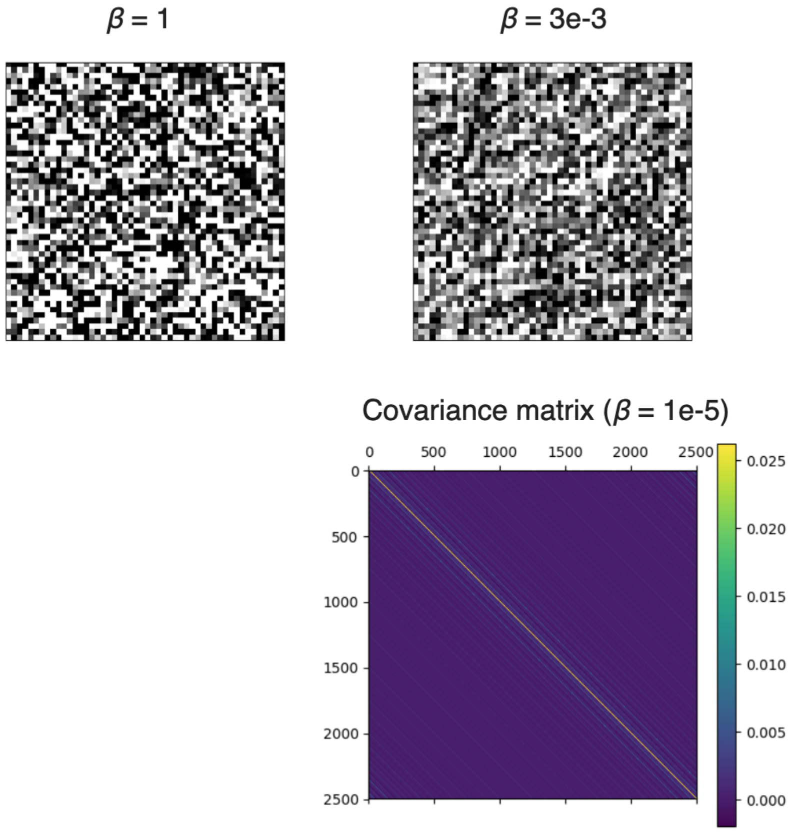


Figure 1-12: Top row: Correlated 50×50 scenes with three different values of β. $\beta = 1$ implies an IID scene, with increasing correlation as β approaches 0. On bottom: the covariance matrix with $\beta = 10^{-5}$. Note that the covariance matrix is 2500×2500, since it describes the covariance of each of the 2500 scene pixels with each other pixel. The pattern of banding that you see depends on how we choose to flatten the 50×50 scene array into a 2500-entry vector; here, the scene is flattened in row-major order.

We distinguish between two different types of noise.

(Ambient noise): This includes noise sources that are independent of the contribution to the measurements due to the scene of interest. That means that the thing causing the noise isn't light coming from the scene; it's light coming from somewhere else (i.e. "glare"). We model it as additive Gaussian with variance W/n, where W denotes the constant net noise power and each pixel absorbs power proportional to its size, giving rise to the $1/n$ factor.

(Shot noise): This includes measurement noise that depends on the contribution due to the scene of interest. This results in additive Gaussian noise of variance $\rho \cdot \frac{J}{n}$ (proportional to the net power of light that goes through the aperture). This noise model comes from the fact that photon detection in reality follows a Poisson distribution; as the number of photons detected becomes large, that distribution will approach a Gaussian distribution with identical mean and variance. Hence, the variance of the number of detected photons (i.e. the strength of the shot noise) will be proportional to the intensity of the received light from the scene.

Ambient noise should be more important in passive non-line-of-sight imaging applications using accidental cameras (in other words, the applications described in Chapter 4) because in that application, the bulk of the light reaching the observation will generally come from sources that aren't the scene of interest, like overhead lighting or sunlight.

Shot noise should be more important in designer-camera applications using coded apertures (in other words, the applications most relevant to the considerations described in Chapter 3) because in those applications, the camera will presumably be designed in such a way as to prevent glare.

Now that we've established the noise model, we can finally write down the equation relating the scene to the observation.

$$y = Ax + \eta \tag{1.2}$$

$$x \sim \mathcal{N}(\mu\mathbf{1}, \mathbf{Q}) \tag{1.3}$$

$$\eta \sim \mathcal{N}(0, (W + \rho \cdot J)/n) \tag{1.4}$$

$\mathbf{Q}$ is the covariance matrix of the scene. $\mu = J/n$ as described earlier. J is the total radiance of the scene, whereas W is a parameter that describes the level of glare (the scene-independent noise).

Chapter 2

Related Work

This thesis studies many aspects of occluder-based imaging, and as such relates to a lot of different past work in different ways. The first and foremost contribution of the thesis is the introduction of a mutual-information theoretic framework in Chapter 3 for comparing different imaging systems, but of course this thesis is not the first to describe such a framework; Section 2.1 briefly describes others. Along the same vein, Section 2.2 describes past work in coded-aperture imaging—which is most relevant, once again, to Chapter 3, because implicit in the study of occluder optimization is the assumption that you are designing your occluder, and can choose it to be whatever you like.

Next, Section 2.3 describes the past work in non-line-of-sight imaging, both active methods and passive. Section 2.4 describes other work that make use of the convolutional nature of occlusion under certain conditions, a concept which is central to the understanding of occlusion in this thesis. Finally, Section 2.5 describes recent work in the field of non-line-of-sight (NLoS) imaging which makes use of machine-learning techniques to improve upon the state of the art. Naturally, all three of these sections relate more closely to Chapter 4, which is about opportunistic occluder-based imaging systems, whose primary purpose is in NLoS imaging.

2.1 Past frameworks for comparing imaging systems

There has been extensive past work studying the benefits of nontraditional cameras, including coded-aperture cameras [82, 114, 40, 41, 66], plenoptic cameras [2, 76] and phase plates [17, 37]. The focus of this thesis is on occluder-based cameras in particular, but the framework is flexible enough to handle a wide variety of different possible imaging systems.

This work is not the first to try to build a common framework with which to compare different imaging systems. Levin et al. [67] was one of the first pieces of work to compare imaging performance across different modalities (where a "modality" here means a lens-based camera, an occluder-based camera, a plenoptic camera, etc.) analytically. The core idea of [67] is to compare the error in the reconstructed 4D light-field, given a 2D observation at the sensor, across each modality. A "light-field" here is defined as a 2D view of the scene, as seen from each point on a 2D sensor array (hence the light field is a $2 \times 2 = 4$ dimensional array). See Figure 3-29.

Naturally, this approach requires a prior. Without any prior (or, equivalently, a uniform prior over all possible light-fields), different, non-degenerate imaging systems will have equivalent performance; all of them will give an equally underconstrained view of the light-field. Thus, that the metric of [67] will give very different results depending on what prior you choose.

Fortunately, [67] also addresses the issue of light-field priors in detail. While past work [46, 85] use a traditional Gaussian (or fat-tailed Gaussian) prior on light-fields, the work of [67] uses a mixture-of-Gaussians prior, which allows them to better model the special structure of most light-fields. Light-fields are unlike ordinary spaces extrapolated into four dimensions, because variance along the "disparity slope" of a light-field is usually much lower than the spatial variance. See Figure 2-1 for a brief explanation of this concept.

This thesis mostly bypasses this issue altogether, as most of the analysis here is restricted to planar reconstructions. This does mean that the framework presented

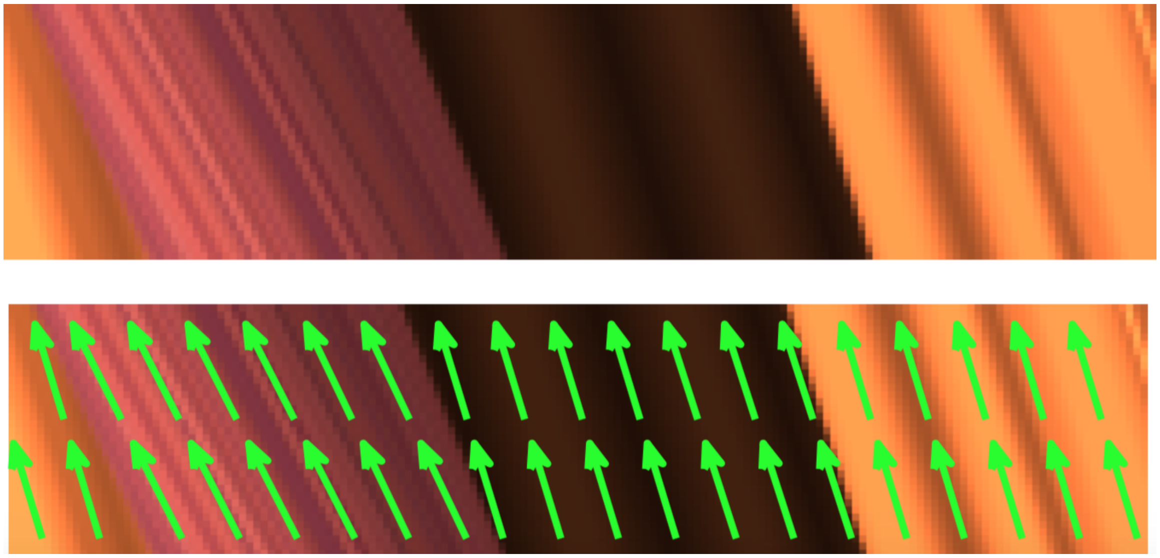


Figure 2-1: This figure is due to the authors of [67]. Top: A 2D light-field of a 1D scene. The horizontal axis indicates space in the scene, while the vertical axis indicates spatial movement along the sensor array. Bottom: the green arrows, inferred by the system of [67], point in the direction of the estimated disparity slope. Along them, the variance of the light-field is very small. This phenomenon is due to the fact that a light-field is simply the scene, seen from every possible sensor on the sensor array; as you move your vantage point along the sensor array, the scene will be nearly the same, except shifted.

in this thesis is likely to be a worse choice that that of [67] if the application under consideration involves depth perception, for example (which is unfortunate, since resolving depth is one of the many applications of coded-aperture cameras, see [66, 114]). The framework proposed by this thesis can be adapted to light-fields, or other situations where depth is relevant (as covered by Sections 3.15 and 3.16), but it's not its primary purpose.

However, by focusing not on light-fields but on planar reconstructions, the framework proposed by this thesis has a few advantages. Firstly, its results are much less dependent on the choice of prior over scenes. In particular, although the results vary somewhat based on how strong the level of scene correlation is, the exact pattern of scene correlation matters very little. Additionally, the results of this thesis are expressed in terms of mutual information rather than mean-squared error (a choice whose implications are explored in Section 3.10). Ordinary photography produces planar reconstructions, and so planar reconstructions were the main focus of this thesis as well.

Another framework for computational imaging was presented by [48]. Unlike this thesis, the focus of their work was on active methods for computational imaging. Moreover, the authors didn't directly compare the different modalities. They did, however, introduce a particular set of notation common to the different methods they studied.

2.2 Coded-aperture imaging

Coded-aperture imaging is the study of imaging using a camera with a known pattern of occlusion, either in combination with a lens, or without any lens at all. Among the earliest and simplest instances of coded-aperture imaging are those based on pinhole structure [41, 111] and pinspeck (anti-pinhole) structure [31], though more complex structure is often used. Other methods involve cameras that uses a mask in addition to a lens to, e.g., facilitate depth estimation [66], deblur out-of-focus elements in an image [114], enable motion deblurring [82], and/or recover 4D lightfields [102]. Some forgo the lens altogether to decrease costs and/or meet physical constraints [38] [9], or for applications that use wavelengths of light unsuitable for a lens, such as nuclear medicine imaging [113]. More generally, a number of rich extensions to the basic methodology have been developed; see, e.g., [18] and the references therein. A focus of this thesis is spectrally flat sequences, also known as uniformly redundant arrays. These arrays have a rich history in computational imaging, both for coded-aperture cameras [47, 40, 41], and in tomography [27, 23].

Perhaps the earliest analysis in this area is that of [111], which studies the resolution limits of pinholes of various sizes. From this perspective, this thesis can be viewed as extending and generalizing such analysis to a broader range of coded aperture systems. The analysis of [111] is easily extended to an NLoS setting [32, 97].

Coded apertures can take on a variety of forms. Although this thesis concerns itself only with square occlusion grids for 2D occluders, hexagonal grids exist which achieve all of the desirable properties of coded apertures [55] (see Figure 2-2). Moreover, although many coded-aperture systems use a mask in addition to a lens, this thesis primarily analyzes systems that use only a mask and no lens (though there is no reason why the framework described in Chap 3 can't be used to analyze systems that include both a lens and a mask).

2.3 Non-line-of-sight imaging

Occluder-based imaging methods, especially methods that make use of accidental occluders, often have non-line-of-sight imaging (or NLoS imaging) as their primary application. Non-line-of-sight imaging is the study of imaging objects that are out of direct view.

Active Methods

Many approaches to NLoS imaging involve a combination of active laser illumination and time-of-flight cameras [63, 88, 107]. These methods, called *active methods*, work by illuminating a point on the visible region that projects light into the hidden scene.

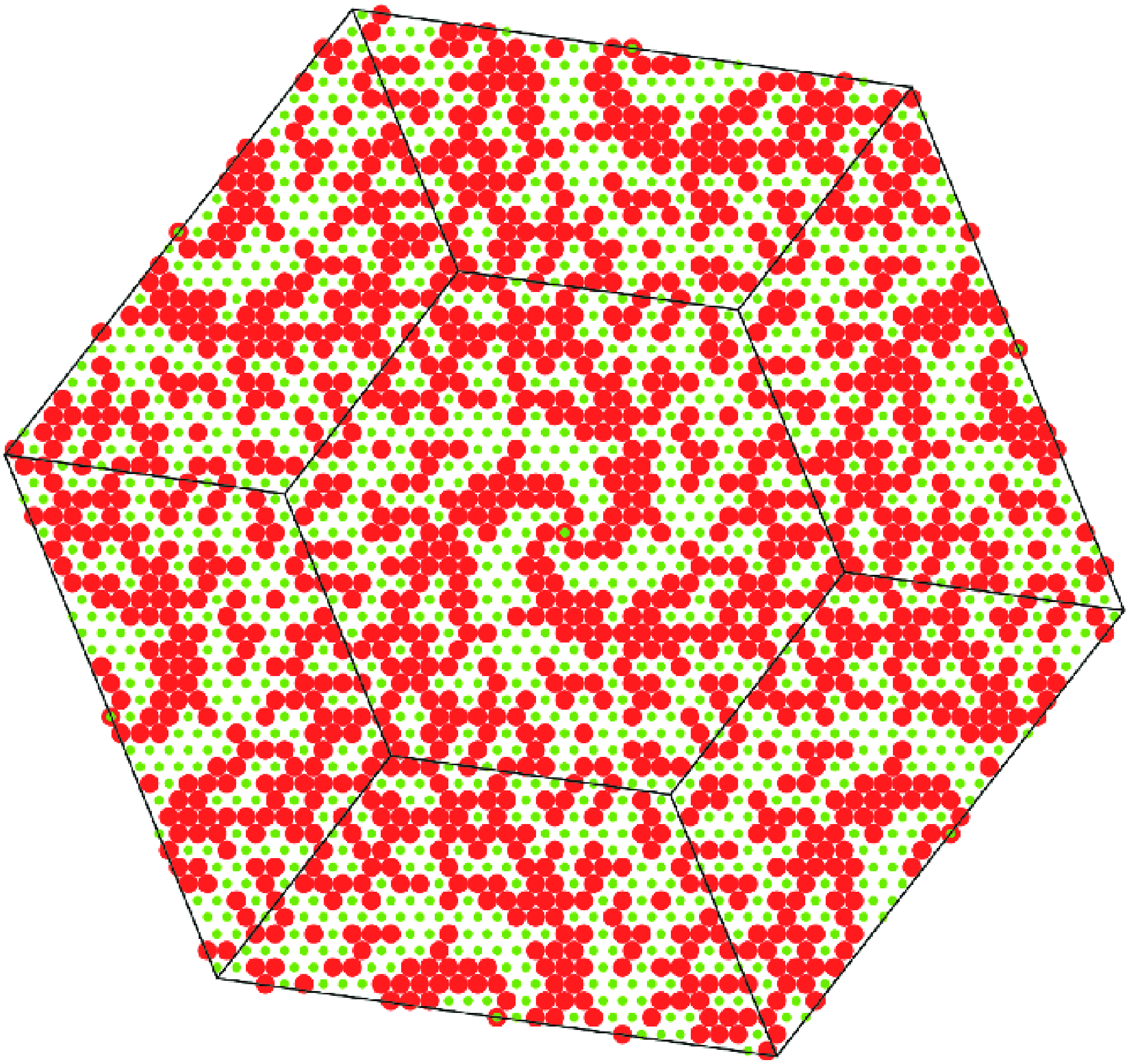


Figure 2-2: A hexagonal grid, due to [55], which achieves spectral flatness, making it potentially desirable for coded-aperture imaging systems.

Then, structure in the scene can be inferred from the time it takes for that light to arrive at detectors [58, 89, 96, 98]. These methods have been used to count hidden people [106], or to infer location, size and motion of objects [43, 57, 78].

Passive Methods

Other recent approaches, called *passive methods*, rely on ambient light from the hidden scene or elsewhere for inference. These approaches range from using naturally-occurring pinholes or pinspecks [32, 97] to using edges [16] to resolve the scene. Passive methods that make use of arbitrary occluders can be divided into calibrated methods [12, 8] and uncalibrated methods [110, 6], with the former algorithms requiring a calibration stage in which careful measurements of the configuration are taken, and the latter not requiring such a stage. In general, complex occluders or imaging elements are easily handled by the former class of methods, but not by the latter class.

Passive methods also sometimes make use of infra-red-spectrum (IR-spectrum) light [70] or even higher-wavelength light [3] to sense emanations created by the scene elements of interest themselves for non-line-of-sight imaging. The advantage provided by IR-based methods over methods using visible light is twofold: firstly, there is often light coming directly from the objects of interest in the scene, rather than as a secondary reflection off of elements in the scene from an overhead light, or the Sun. Secondly, higher-wavelength-light reflections are closer to being specular than Lambertian on a matte wall, relative to visible-light reflections. (See Section 1.2 for more explanation of what's meant by this.) However, longer-wavelength light limits the resolution of potential reconstructions. This gives shorter-wavelength light more potential to yield high-resolution reconstructions; it should be admitted, however, that the resolution of state-of-the-art passive NLoS reconstructions using visible light remains far below that potential [110, 6, 16, 97].

2.4 Convolutional models of occlusion

Accurate modeling of occlusion and light transport is a core problem in computer graphics [112, 19, 105, 5]. One imporant difference between the study of occlusion in the field of computer graphics and the field of computational imaging is that the former is more interested in forward models (i.e. understanding what the observation will look like, given the objects in the scene), whereas the latter is generally interested both in forward and backward models (backward models meaning inferring the objects in the scene from the observation).

Though the one major issue in graphical applications is hierarchical culling, i.e. knowing which occluding layer to query in order to correctly display the environment [13, 35], which is not a problem of as much interest in the field of computational imaging, there is nevertheless important overlap between the two fields. In particular, a convolutional model of occlusion is common to both fields [91, 10, 109].

Recently, it was shown by that Ahn et al. [4] that active, time-of-flight methods can also be modeled as using a convolutional operator in their forward model, leading to computational savings.

2.5 Computational imaging using machine learning

Chapter 4 describes a variety of non-line-of-sight imaging methods. None of the methods described, save for that in [6], use machine learning. As machine vision techniques become increasingly powerful, however, machine-learning techniques are becoming more popular as a means to automatically infer the forward and backward models. They are of particular interest for blind problems, such as the blind occluder-based recovery problem of [110] or the blind clutter-based recovery problem of [6].

The work of [6] uses the Deep Image Prior, a concept originally due to Ulyanov et al. [100]. The "Deep Image Prior" is the remarkable discovery that due to their structure, convolutional neural nets inherently impose a natural-image-like prior on the outputs they generate, even when they are initialized with random weights and without any pretraining. Since the publication of [100], there have been several other papers that make use of the Deep Image Prior for a variety of applications, including compressed sensing [101], image decomposition [42], denoising [28], and image compression [54].

Aittala et al. [6] use the Deep Image Prior to characterize the prior over transfer matrices and over scenes—somewhat analogously to the work of Levin et al. [67], because the core idea is introduced in order to better describe an object for which an ordinary Gaussian spatial prior would be inadequate (in the former case a transfer matrix, in the latter case a light field).

Other recent non-line-of-sight methods to incorporate general-purpose learning algorithms are [99], which uses a Monte Carlo rendering phase first to estimate facet locations, followed by a stochastic gradient descent (SGD) phase to synthesize those estimates. Another such result is [61], which uses an iterative backprojection algorithm to improve reconstructed estimates of 3D scenes.

Chapter 3

Analysis and Optimization of Aperture Frames

The bulk of this chapter is concerned with which aperture frames (and which corresponding transfer matrices) are best under which conditions; "best," of course, requires a metric to optimize over, and throughout most of this chapter I use maximizing mutual information as this metric. Section 3.8 discusses minimizing mean-squared error as a possible metric instead, and within that section, Subsection 3.10 explicitly compares and contrasts the two metrics. (The one-sentence summary of Subsection 3.10 is that switching between these two metrics doesn't change much.)

3.1 Mutual Information

Suppose that as described previously, we have:

$$y = \widetilde{\mathbf{A}}x + \eta \tag{3.1}$$

$$x \sim \mathcal{N}(\mu\mathbf{1}, \mathbf{Q}) \tag{3.2}$$

$$\eta \sim \mathcal{N}(0, (W + \rho \cdot J)/n) \tag{3.3}$$

Here, $\widetilde{\mathbf{A}}$ is the transfer matrix, x and y are the scene and observation, respectively,

$\mathbf{Q}$ is the covariance matrix of the scene, W and J are constants parametrizing the strength of the ambient and shot noise, respectively, and ρ is the transmissivity of the aperture frame. (Recall that the ambient noise is noise due to nuisance light sources, like glare from the Sun on your computer screen, and shot noise is noise due to illumination from the scene.)

Then, the mutual information (MI) between the measurements $y_j, j \in [n]$ and the unknowns $f_i, i \in [1, n]$ of the imaging problem for model 3.3 is:

$$\mathcal{I} = \log\det\left(\frac{1}{\sigma^2}\widetilde{\mathbf{A}}\mathbf{Q}\widetilde{\mathbf{A}}^T + \mathbf{I}\right),$$

where the noise variance $\sigma^2 = (W + \rho \cdot J)/n$.

We'll define $A = n\widetilde{\mathbf{A}}$, so we can equivalently write:

$$\mathcal{I} = \log\det\left(\frac{1}{n^2\sigma^2}A\mathbf{Q}A^T + \mathbf{I}\right)$$

3.2 High SNR, IID scene, Occluding mask

In this section, I will go into great detail about a regime that appears unphysical, but gives us a lot of insight into a variety of different important regimes. It also has important mathematical implications.

Consider the mutual information equation from the previous section:

$\mathcal{I} = \log\det\left(\frac{1}{\sigma^2}\widetilde{\mathbf{A}}\mathbf{Q}\widetilde{\mathbf{A}}^T + \mathbf{I}\right)$

Suppose we make the following assumptions:

1. The scene is IID, $\mathbf{Q} = kI$.

2. The SNR is very high, $\sigma \ll k$, so we can ignore the identity term.

3. The aperture frame only occludes light or lets it through; it doesn't redirect the light.

4. The aperture frame is halfway in between the scene and observation plane.

5. The scene, aperture frame, and occluder are all planar.

6. The observation plane is diffuse.

7. The paraxial approximation applies.

Suppose now that we want to find the best possible aperture frame under these conditions, i.e. we want to find the best possible transfer matrix that satisfies the constraints above. Given the third point, we may call this aperture frame an "occluder," since it only occludes or doesn't occlude light. What does this transfer matrix look like? What are its elements? Each element $\widetilde{\mathbf{A}}_{ij}$ corresponds to the amount of light a little piece of the scene x_i will contribute to a little piece of the scene y_j. Obviously, when the occluder blocks the light from getting from x_i to y_j, $\widetilde{\mathbf{A}}_{ij} = 0$. What about when that doesn't happen? We can look at Section 1.2 to remind ourselves of what the Lambertian BRDF is. Because we are using the paraxial approximation (see our last assumption), we'll have:

$$\widetilde{\mathbf{A}}_{ij} = \frac{1}{2\pi d n} \tag{3.4}$$

in flatland, and:

$$\widetilde{\mathbf{A}}_{ij} = \frac{1}{4\pi d^2 n} \tag{3.5}$$

in a 3D world. Here, d is the distance between scene and observation. The important thing about Formulas 3.4 and 3.5 is their dependence on $1/n$, which is due to the fact that as we discretize more and more finely, each little piece of the observation will get smaller and smaller (and therefore get proportionally less light). In fact, if we ignore the whole rest of the formula, then we get $\widetilde{\mathbf{A}}_{ij} = 1/n$. Doing so ignores a constant; while it will cause us to get the wrong mutual information in the real world (by an amount of $2n \log\left(\frac{1}{4\pi d^2}\right)$), it won't change the relative ranking of which occluders are best, which is what we're really interested in. And if we're willing to ignore that constant—or subsume it into σ, which is also a constant—then something

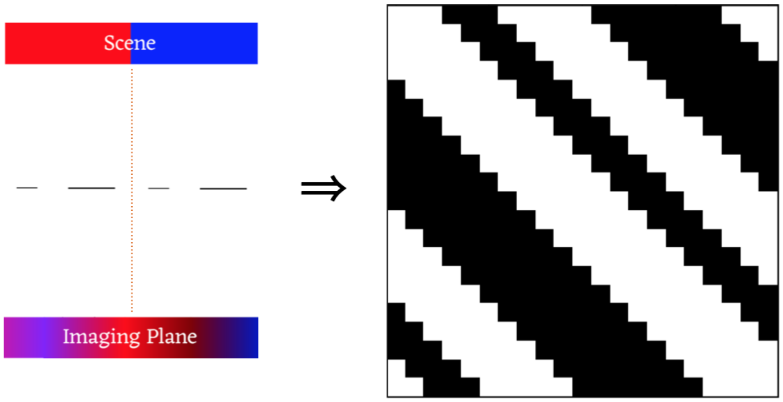


Figure 3-1: A once-repeating occluder, with its associated transfer matrix. As is apparent, the transfer matrix is circulant, with each row of the transfer matrix being a rotation of the pattern that is repeated once. See Figure 1-9 for a reminder on how the transfer matrix is created from an aperture frame.

nice happens: given that $\widetilde{\mathbf{A}}$'s entries will all be either $1/n$ or 0, A will be a binary matrix! (A binary matrix is one whose entries are all either 1 or 0, and $A = n\widetilde{\mathbf{A}}$.) This lets us express the mutual information in terms of a binary matrix, which is tremendously convenient analytically; we'll be using this fact a lot in the sections that immediately follow.

Now, what transfer matrix maximizes the mutual information? Naturally, it will be whichever transfer matrix maximizes $\log\det\left(\frac{1}{\sigma^2}A\mathbf{Q}A^T/n^2 + \mathbf{I}\right)$, which given our assumptions is equivalent to maximizing the determinant of A^TA. This corresponds to maximizing the product of the norms of the transfer matrix's singular values, since each eigenvalue of A^TA corresponds to the norm squared of one of the eigenvalues of A.

If we further assume that A is circulant, not just Toeplitz, then that tells us that in fact the singular values of $\widetilde{\mathbf{A}}$ have the same norm. This is a convenient assumption that doesn't necessarily match reality. Under which physical conditions will the transfer matrix actually be circulant rather than Toeplitz? This is somewhat unintuitive, but it corresponds to the scenario in which the occluding pattern of the aperture frame repeats itself once. Once is enough! See Figure 3-1.

Assuming that the transfer matrix is circulant, not just Toeplitz, is tremendously convenient. It means that we can compute the mutual information between the scene and the observation extremely efficiently, since the eigenvalues λ_i are given by the Fourier transform of the first row of the transfer matrix, for which the time to compute is log-linear in n (as opposed to $O(n^3)$ for a general determinant). And since $|\lambda_i| = |\sigma_i|$, that gives us all the information we need to compute $\log\det(AA^T)$.

But realistically speaking, can we restrict ourselves just to circulant transfer matrices rather than Toeplitz ones? After all, occluders that give rise to that kind of transfer matrix make up only a very small fraction of all possible occluders. The answer is that it depends on what you want, but if you're only concerned with finding the *best* possible occluder, given all the assumptions we made previously, restricting your attention to circulant transfer matrices costs you very little. The following section explains why that is.

3.2.1 Hadamard's Bound

The purpose of this section is to justify the following bound on all $\{0,1\}$ matrices B. This material also appears in [20].

$$|\det B| \leqslant 2^{-n}(n+1)^{(n+1)/2}, \tag{3.6}$$

In this section, by a *binary* matrix we mean a matrix whose elements are in one of the sets $S_{01} := \{0,1\}$ or $S_{\pm 1} := \{-1,1\}$. It will be clear from the context which of these two cases is being considered. A *binary circulant* is a circulant matrix whose elements are in S_{01} or $S_{\pm 1}$.

There is a natural correspondence between the integers $\{0,1,\ldots 2^n-1\}$ and the binary circulant matrices of order n. If $N \in \{0,1,\ldots,2^n-1\}$ has the representation

$$N = \sum_{j=0}^{n-1} 2^{n-1-j} b_j,$$

so may be written in binary as $b_0 \ldots b_{n-1}$, we associate N with $\mathrm{circ}(a_0,\ldots,a_{n-1})$, where $a_j = b_j$ in the case of S_{01}, and $a_j = 2b_j - 1$ in the case of $S_{\pm 1}$.

The *maximal determinant problem* is concerned with the maximal value of $|\det A|$ for an $n \times n$ binary matrix A. The *Hadamard bound* [49] states that, in the case of binary matrices A over $\{\pm 1\}$, we have

$$|\det A| \leqslant n^{n/2}. \tag{3.7}$$

Moreover, Hadamard's inequality is sharp for infinitely many n, for example powers of two or n of the form $q+1$ where q is a prime power and $q \equiv 3 \pmod 4$ (Paley [77]).

There is a well-known connection between the determinants of $\{0,1\}$-matrices of order n and $\{\pm 1\}$-matrices of order $n+1$. This implies that an $(n+1) \times (n+1)$ $\{\pm 1\}$-matrix always has determinant divisible by 2^n. See [75] for details. We give an example with $n = 3$, starting with an $n \times n$ binary matrix B and ending with an $(n+1) \times (n+1)$ $\{\pm 1\}$-matrix A, with $\det A = 2^n \det(B)$.

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \xrightarrow{\text{double}} \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix}$$

$$\xrightarrow{\text{border}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 0 & 2 & 2 \end{pmatrix} \xrightarrow[\text{first row}]{\text{subtract}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 \end{pmatrix} = A.$$

The doubling step is the only step where the determinant changes, and there it is multiplied by 2^n.

Thus, Hadamard's bound (3.7) gives the bound

$$|\det B| = 2^{-n}|\det A| \leqslant 2^{-n}(n+1)^{(n+1)/2}, \tag{3.8}$$

which applies for all $\{0,1\}$-matrices B of order n. We shall refer to both (3.7) and (3.8) as *Hadamard's inequality*, since it will be clear from the context which inequality is intended.[1]

3.2.2 Binary Circulants Achieving Hadamard's Bound

Now we get to the reason that, if you're only concerned with finding the *best* possible occluder, restricting your attention to circulant transfer matrices costs you very little. The reason is that there are several constructions for binary circulant matrices that achieve Hadamard's bound—this *despite* the fact that Hadamard's bound is general for all binary matrices, not just circulant ones! It is very lucky that among the binary circulant matrices, which comprise a tiny subset of all binary matrices, that some of those binary circulant matrices achieve determinants as large as any from *all* binary matrices of the same size. I'll call these remarkable binary circulants "determinant-maximizing binary circulants," or DMBCs.

There are four known constructions for DMBCs. All four of these constructions yield a matrix whose eigenvalues are all equal, save for the first; the first eigenvalue has a value of $(n+1)/2$, and every other eigenvalue has a value of $\sqrt{n+1}/2$. The dot product of any pair of rows in a DMBC from one of these four constructions is $(n+1)/4$. In each construction, there are $(n+1)/2$ elements with value 1 and $(n-1)/2$ elements with value -1. It follows from that last sentence that if you take one of these constructions and replace all the 0 entries with -1 entries to yield a $\{1,-1\}$ matrix, the dot product of any pair of rows will be -1.

Thanks to this last point, DMBCs have a close relationship to Hadamard matrices. Hadamard matrices are $\{1,-1\}$ matrices (not necessarily circulant) for which every pair of rows is orthogonal, that is, the dot product of any pair of rows is 0. (The same is true of pairs of columns.) Any size-n DMBC from one of these four constructions can be easily adapted to create a size-$(n+1)$ Hadamard matrix as follows: replace all the 0 entries with -1 entries in the DMBC, and then add a first row and a first column of all -1 entries. The dot product of the first row with any other row will be 0 (because every other but the first is exactly half 1 entries and half -1 entries). The

[1] In fact, Hadamard in [49] proved a more general inequality than (3.7), and as far as we are aware he never stated (3.8) explicitly. A simple proof of (3.7) is given by Cameron [21].

dot product of every row other than the first with any other row other than the first will also be 0 (because the dot product of each pair of rows before adding the extra row and column was -1, and then adding the extra column adds an extra 1 to the dot product). Hence each size-n DMBC yields a size-$(n+1)$ Hadamard matrix.

Hadamard matrices of this kind are known in the literature as "Hadamard matrices with circulant core." Here is more detail about each of the known constructions:

Theorem 1 (Hadamard circulant core construction)**.** *A Hadamard matrix of order $n+1$ with circulant core of order n exists if*

(1) $n \equiv 3 \pmod 4$ *is a prime;*

(2) $n = p(p+2)$, *where p and $p+2$ are prime;*

(3) $n = 2^k - 1$, *where k is a positive integer; or*

(4) $n = 4k^2 + 27$, *where k is a positive integer and n is a prime.*

Proof. Case (1) is due to Paley [77]; case (2) is due to Stanton and Sprott [92] and also Whiteman [104]; case (3) is due to Singer [90]; and case (4) is due to Hall [50, Theorem 2.2]. □

Hall [50, p. 980] remarks that case (4) is subsumed by case (1), since $4k^2 + 27 \equiv 3 \pmod 4$, but we mention case (4) since Hall's construction is different from that of Paley.

We do not know if the list given by Theorem 1 is exhaustive. The exhaustive search given in [20] shows that for $1 \leqslant n \leqslant 52$, only those n given by Theorem 1 can provide a Hadamard matrix of order $n+1$ with a circulant core. Also, a circulant $\{0,1\}$-matrix of order $n \leqslant 52$ can achieve the upper bound (3.8) if and only if $n \leqslant 4$ or n satisfies condition (1), (2) or (3) of Theorem 1.

This gives us the four known constructions for DMBCs. Note that the fourth construction is completely redundant with the first, since any prime n such that $n = 4k^2 + 27$ where k is a positive integer is guaranteed to also be prime and congruent to 3 mod 4! It is only considered a separate construction because it yields an additional DMBC beyond the one given by the first construction (and isn't a trivial transformation). The fourth construction is therefore of no additional practical value to us: we can't use it to create a mask that images at any level of resolution not already available to us.

It should be noted that sequences from all four of these constructions have a "flat spectrum," meaning that the magnitudes of all of the frequencies of each sequence have a value of $\sqrt{n+1}/2$ save the first, which has a magnitude of $(n+1)/2$. This means that these four constructions have identical performance; they differ only in what values of n they exist for. These sequences, which are sometimes called "spectrally-flat sequences" or "uniformly redundant arrays," have a rich history in coded-aperture imaging [41, 47, 40] and tomography [23, 27, 23]. It's well-known that these sequences are good for coded-aperture imaging; moreover, it's reassuring that they come out on top in our mutual-information-based framework (provably under the simplified conditions of this section, empirically under the more complex conditions of later sections).

It's worth giving more detail about the first construction, since it's by far the most common one over the real numbers; there are many more primes congruent to 3 mod 4 than there are powers of 2 or products of twin primes! Indeed, it's common enough that no matter what level of resolution you need, there will be a reasonably suitable mask at a nearby resolution level, thanks to the first construction.

The first construction, due to Paley, is as follows:

If n is prime and congruent to 3 mod 4:

$$x_i = \begin{cases} 1 & \text{if } \left(\frac{i}{n}\right) = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, $\left(\frac{i}{n}\right)$ is the Legendre symbol. It's equal to 0 if i is a multiple of n, and is otherwise equal to 1 if i is a quadratic residue (meaning a perfect square) modulo p and -1 if not. It's very easily computed, since by Euler's criterion, we have, for any a and any prime p:

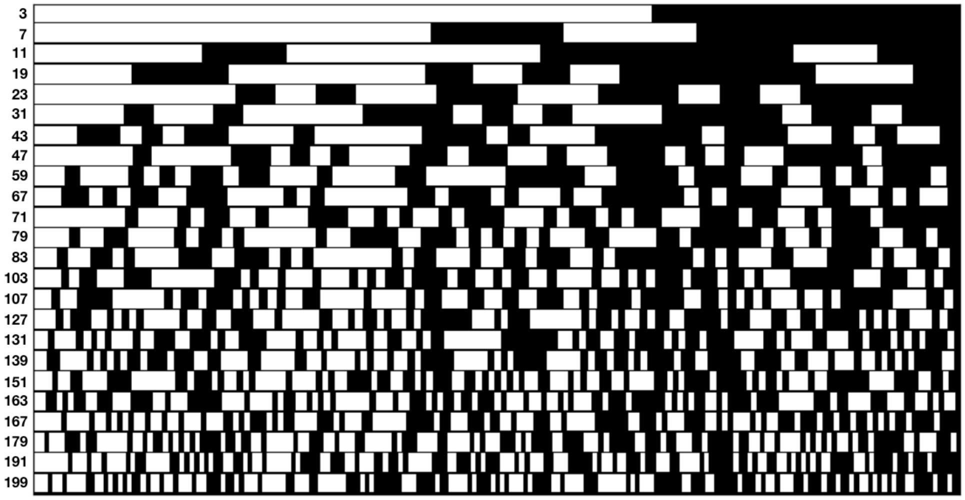


Figure 3-2: These are the first few Paley sequences, i.e. on-off patterns whose spectrum is flat and that yield a DMBC (determinant-maximizing binary circulant) when used as the first row of a circulant matrix. On the left is n, the number of on/off patches used. Note that for a Paley sequence to exist, n must be prime and congruent to 3 mod 4.

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \mod p$$

Figure 3-2 shows the first few sequences of the Paley construction. As you can see, it has a very “random-looking” appearance. Of course, the construction is very much non-random; what gives it that appearance, though, is its non-self-repeating character. All four constructions, in fact, repeat themselves as little as possible. That’s not surprising from the point of view of wanting to keep the sequence’s Fourier spectrum flat, of course. But the way I like to think of it from an imaging point of view is that these sequences try to make the different possible shadows cast by a light source in a variety of different locations as different as possible from each other. A light source at each possible point in the (implicitly planar) scene will yield a different rotation of the occluding sequence, so the best sequences for distinguishing light sources at different locations will be sequences that are orthogonal to their own rotations. See Figure 3-3 for an explanation of this phenomenon.

In any event, because DMBCs achieve the maximal possible mutual information

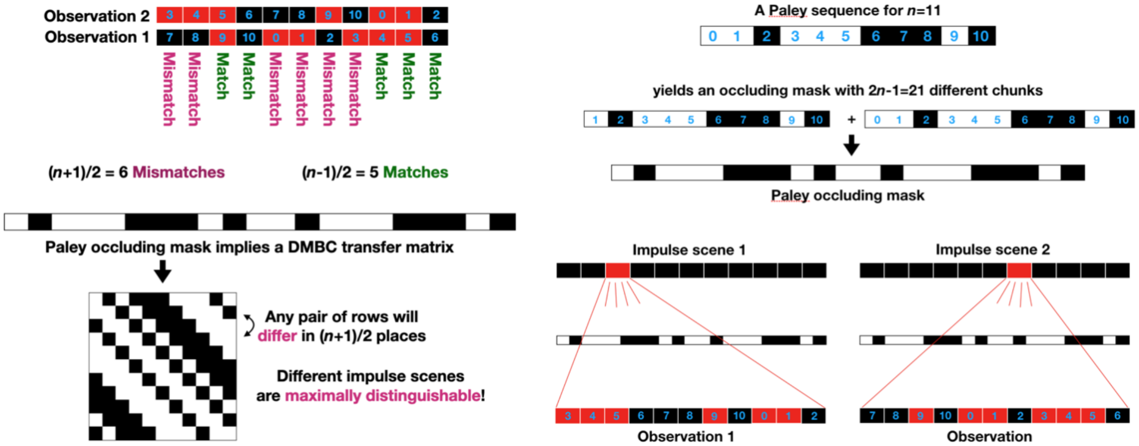


Figure 3-3: Paley sequences, the "optimal" occluding masks that be constructed from them, and the transfer matrices associated with those masks. Left: we consider the Paley construction applied to the case $n = 11$. The Paley construction gives us a binary sequence, with 1's (white, or "on") element i of the sequence if i is 0 or a quadratic residue modulo n, and 0's (black, or "off") for elements i of the sequence if i is not a quadratic residue modulo n. From this sequence we get an occluding mask, which consists of the sequence twice; every element is repeated exactly once except for the 0 element, which is in the center. Given the paraxial approximation and assuming the occluder is halfway in between observation and scene, an impulse scene will cast a shadow that corresponds to half the occluder, which is some rotation of the original Paley sequence. Right: sequences given by the Paley construction (as well as any other flat sequence) are as different as possible from their own rotations; this means that depending on where the impulse light source is in the scene, the cast shadows will be as different as possible. This makes reconstructing the location of a point light source, given a cast shadow, as easy as possible.

of any binary matrix of their size, the fact that we restricted our attention to circulant matrices rather than considering all possible Toeplitz matrices doesn't matter, assuming the value of n we're using admits the existence of a DMBC. We therefore know that the once-repeating occluder suggested by that DMBC outperforms all other possible occluding aperture frames, including ones that don't repeat themslves.

3.3 Maximizing binary determinants for all values of n

The problem of maximizing the determinant of a binary circulant matrix for general n is much harder. In Brent and Yedidia [20], a method is described for finding maximal determinants of binary circulant matrices at all values of n, but the method described is essentially brute-force, with enough optimizations to make it substantially faster than a naive algorithm, but nothing that changes the exponential base of its runtime.

One might expect a greedy algorithm for optimizing binary sequences to perform well in practice as n gets large. In fact, that is not the case. Figure 3-4 compares the performance of four metrics: the best possible performance as given by Hadamard's upper bound (see Equation 3.8), the best possible performance at that value, as verified by an exhaustive search (as detailed in *Computation of Maximal Determinants*), the performance of the outcome of a greedy search, and the performance of a binary sequence chosen uniformly at random.

For completeness' sake, I'll briefly describe the greedy search algorithm used here. The algorithm is simple: check a bit in the sequence, chosen uniformly at random. If flipping that bit would improve the performance of the binary sequence, the bit is flipped, otherwise it isn't; this process is repeated until $3n$ consecutive steps go by in which no bit is found which would improve the sequence. This algorithm is analogous to the Metropolis-Hastings algorithm, without rejecting candidate moves.

It is plain from Figure 3-4 that the greedy algorithm does not keep pace with the optimal possible performance, even asymptotically—in fact, it's unclear if even the best possible sequences at each value of n without a known DMBC asymptotically keep pace with Hadamard's upper bound. This is a disappointing result. The failure of greedy algorithms to do nearly as well as what is possible has to do with how jagged the space being searched is; the space of determinants of binary circulants is full of local optima, most of which fall well short of the global optimum.

To make this more concrete, I introduce here the concept of a *satellite*. A satellite of a local optimum is a point in a space that will lead to that local optimum if a greedy search is started with the satellite as the starting point. We can closely estimate how many local optima exist by performing many greedy searches to estimate how many satellites per local optimum there are in the space. The more satellites each local optimum has, the fewer local optima must exist. Because there can be only one local optimum per satellite, in fact, if there are $f(n)$ satellites per local optimum and $g(n)$ total vertices (in this case, binary sequences) in the space to be searched, that implies that there are $g(n)/f(n)$ local optima—and that an algorithm that uses repeated greedy searches, initialized at random locations in the space, will take $O(g(n)/f(n))$ time, multiplied by the length of a search. (In this case, greedy searches empirically take linear steps and quadratic time, though it's hard to prove that this is guaranteed.)

Using sampling for small values of n, it's possible to find out how many satellites per local optimum there are in this space. In this case, it seems that the number of satellites grows at a pace of about 1.22^n, implying that a greedy-search-based algorithm should be able to achieve an asymptotic runtime of $\tilde{O}(1.78^n)$ (meaning that the base of the exponent is 1.78, and ignoring subexponential terms)—slightly better than the brute-force algorithm, but not much. See Figure 3-5.

3.4 Maximizing the mutual information of correlated scenes

One useful concept that in the sections that follow is the "effective resolution" (or "effective pixel count") of an imaging system. Given an SNR and a probability dis-

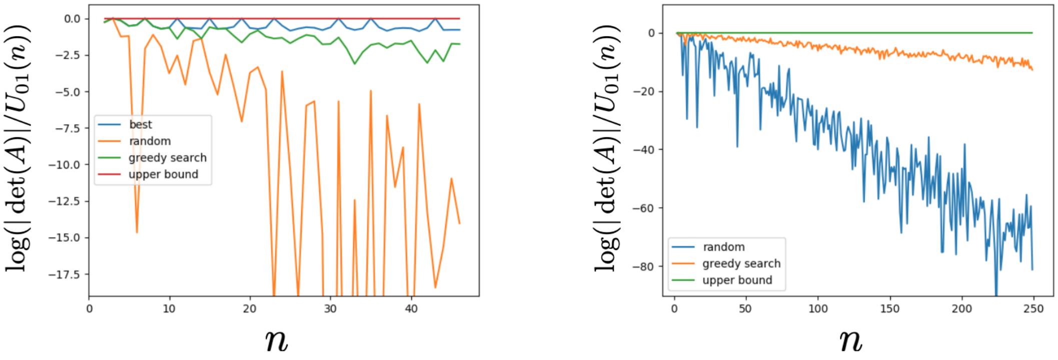


Figure 3-4: Two plots of the ratio of the performance of various sequences to Hadamard's upper bound, $U_{01}(n) = 2^{-n}(n+1)^{(n+1)/2}$. The left plot includes the best possible binary matrix for each value of n (note that the best possible binary matrix achieves Hadamard's upper bound when a known construction for a DMBC exists, and not otherwise). The right plot extends the plot up to higher values of n, but can't include the performance of the best possible binary matrix, because it takes too long to find for values of n much greater than 50.

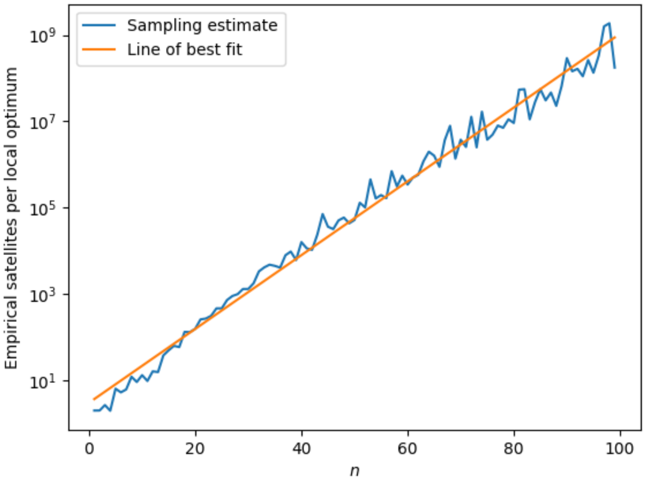


Figure 3-5: The number of satellites per local optimum, as estimated by sampling, including a line of best fit. This suggests approximately 1.22^n satellites per local optimum in the space of maximal determinants of binary sequences, implying a simple search algorithm should be able to find the global optimum in $O(n^2 1.78^n)$ time.

tribution over scenes, the effective resolution will be the best possible reconstruction resolution provided by any occluding aperture frame. We can approximate the effective resolution of an imaging system by considering which scale of spectrally-flat occluder will perform best; see Figure 3-6. This concept is related to, but not the same as, the Nyquist rate (which describes the minimum rate at which a finite bandwidth signal needs to be sampled to retain all of the information). The effective resolution of an imaging system is different because it takes the entire imaging system into account; for example, considering where the aperture frame is positioned relative to the scene and observation plane (as we will do in a later section) will change the effective resolution of the imaging system.

Figure 3-6 is a plot of the approximate number of effective pixels in the system, as a function of the SNR and β (i.e. the level of correlation in the scene).

3.5 Using the best URA as a proxy for the best occluder

As an aside, the sort of plot shown in Figure 3-6 is one I will show a lot, in that section as well as future ones. While the best occluder chosen from among only spectrally-flat masks at various scales is of course not guaranteed to be the best occluder overall, it does a surprisingly close impression of the best occluder overall, as well as giving an easily-digestible sense of the effective resolution of the reconstruction that you can expect.

How close does the best spectrally-flat mask come to the overall occluder in performance? Because we can only estimate the optimal occluder for small values of n, I've duplicated Figure 3-6 but with $n = 15$ in Figure 3-7. Figure 3-7 additionally compares the performance of the best spectrally-flat occluder from among the four shown with the performance of the actual best occluder under each set of conditions, found by exhaustive search. As can be seen, the difference in performance is non-existent for both the low-SNR/high-correlation and the high-SNR/low-correlation regimes,

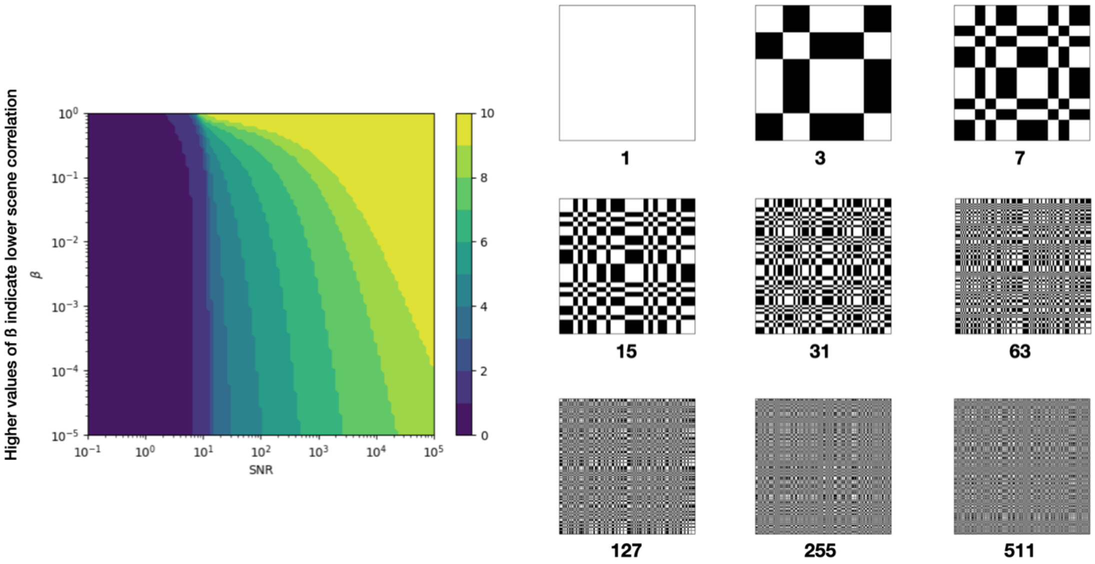


Figure 3-6: Left: the approximate effective pixels-per-side count of scenes generated with a given level of correlation (β) and under a given signal-to-noise ratio (SNR). $n = 1023$. As expected, more correlated scenes and noisier scenes both have fewer effective pixels. Note, however, that even highly correlated scenes can have high effective pixel counts if the SNR is high enough, but if the SNR is low enough the scene will always have a low effective pixel count. Effective pixel count was estimated by choosing which of the nine spectrally-flat masks shown on bottom yielded the highest mutual information. Right: masks corresponding to each of the effective scene pixel counts. Note that these masks repeat themselves once in each dimension, so each mask is $2n - 1 \times 2n - 1$ if the effective pixel count is n. This is due to the phenomenon described in Figure 3-3.

since the empty occluder and the finest spectrally-flat occluder are optimal in each of those regimes, respectively. Only in the intermediate regimes is one of the four spectrally-flat occluders not optimal, and even in those, the difference between those and the true optimal occluder is miniscule (note the scale of the difference plot!).

For larger values of n ($n = 127$), we can compare the performance of picking the best spectrally-flat occluder from among 7 spectrally-flat occluders at different scales to the performance achieved by an occluder found via a greedy search over occluders. See Figure 3-8.

3.6 Reconstructing with a prior

The problem of reconstructing a scene from an observation, given a known scene covariance matrix Q and a known SNR term σ, has a simple MSE solution:

$$\hat{x} = \sigma(\sigma A^T Q A + I)^{-1} A^T y,$$

where y is the observation, $\hat{x}$ is the reconstructed scene, A is the known transfer matrix, Q is the known covariance matrix, and σ is the known SNR. In practice, Q and σ may not be known; they can be guessed, either through trial and error (to produce the best-looking reconstruction) or estimated.

Figure 3-9 shows a comparison with and without a non-identity covariance matrix Q, which I will also refer to as inverting with or without a prior, respectively. As can be seen, the effect of including a non-identity Q ($\beta = 0.1$, using the exponential-decay prior explained in Section 1.4.2) is to reduce the spatial variation in the reconstruction.

3.7 Results from *Analysis and Optimization of Aperture Design in Computational Imaging*

The following sections use material from [109]. They are meant to summarize the results of that paper in a way that will be familiar to a reader of this thesis.

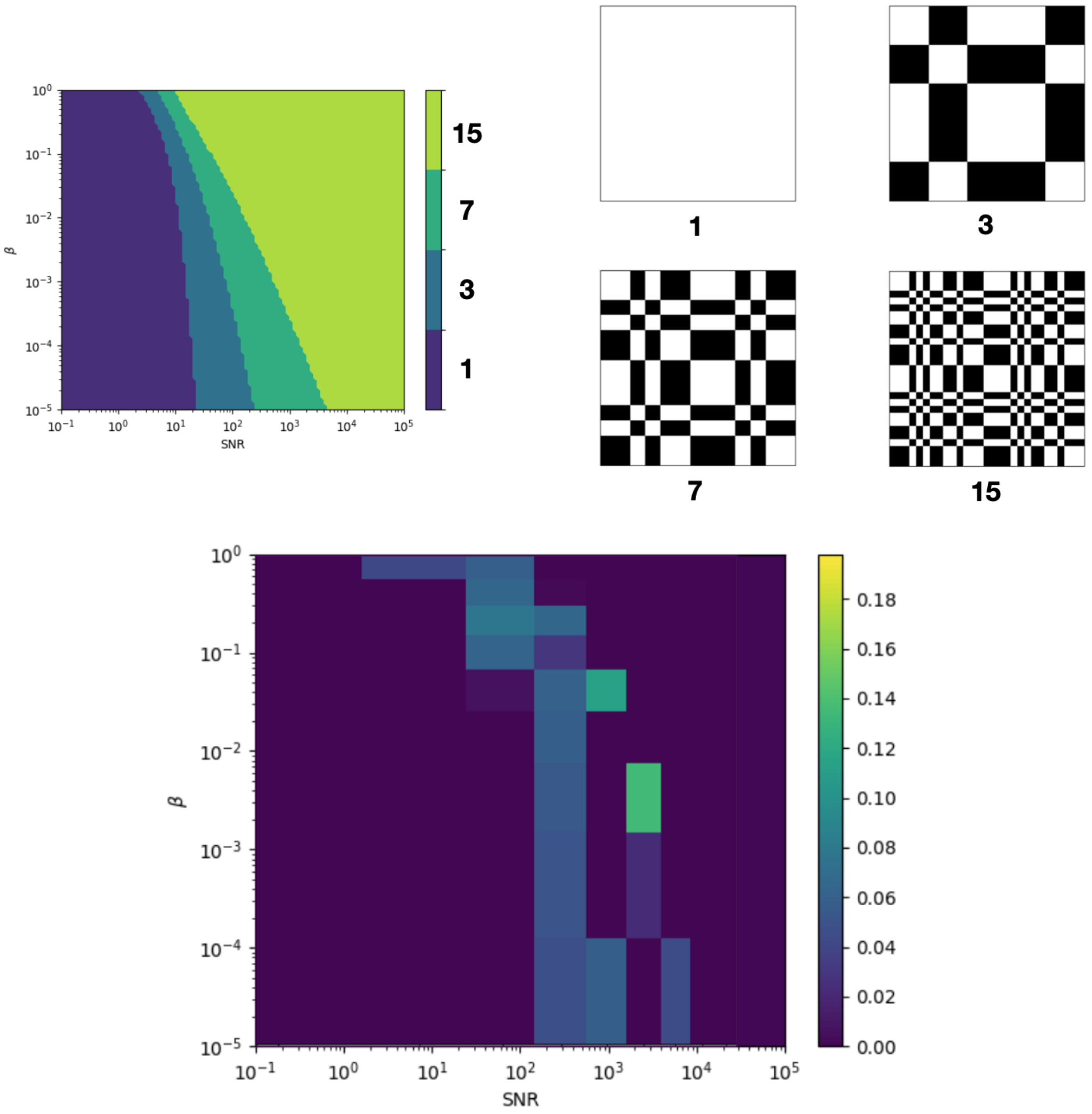


Figure 3-7: Top left: the approximate effective pixels-per-side count of scenes generated with a given level of correlation (β) and under a given signal-to-noise ratio (SNR). Alternately, the plot shows which of the masks shown in the top right is optimal, assuming you are forced to choose one of those four. Top right: masks corresponding to each of the effective scene pixel counts. Bottom: the difference between the MI performance of the best mask from among the four masks in the top right and the best possible mask, as a percentage of the MI performance of the best possible mask, as found by exhaustive search. Thus for $n = 15$, choosing your mask from only the four masks in the upper right is no more than 0.2% worse than using the very best mask.

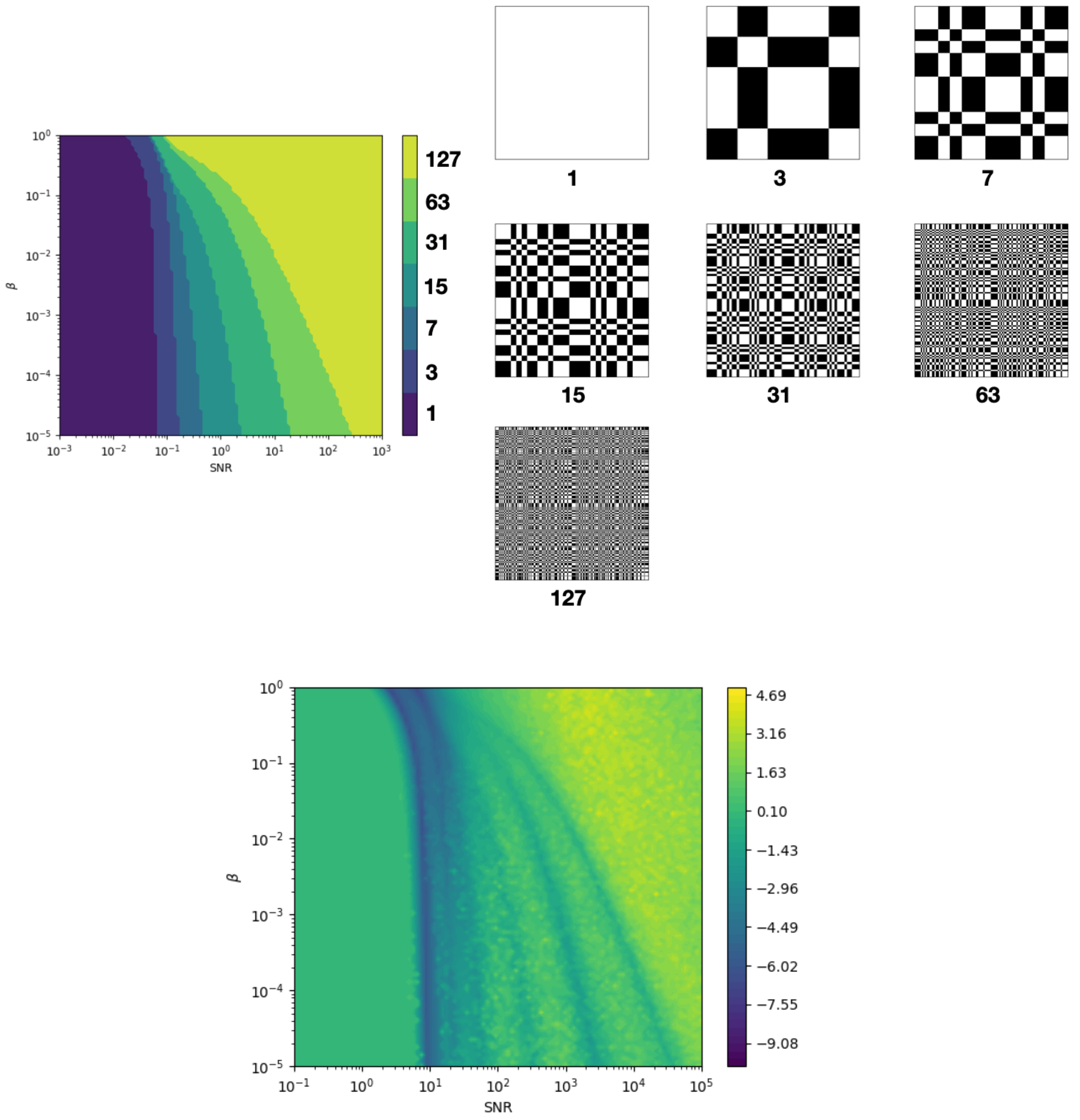


Figure 3-8: Top left: the approximate effective pixels-per-side count of scenes generated with a given level of correlation (β) and under a given signal-to-noise ratio (SNR). Alternately, the plot shows which of the masks shown in the top right is optimal, assuming you are forced to choose one of those four. Top right: masks corresponding to each of the effective scene pixel counts. Bottom: the difference between the MI performance of the best mask from among the seven masks in the top right and the best mask found with a greedy search, as a percentage of the best mask from among the seven masks in the top right. Positive percentages mean that choosing the best mask from a menu is better; negative percentages mean that a greedy search is better. We can see that choosing from a menu of URAs at different scales outperforms greedy search in the high-SNR regime, and does worst at the *boundary* between masks at different scales, which is intuitive.

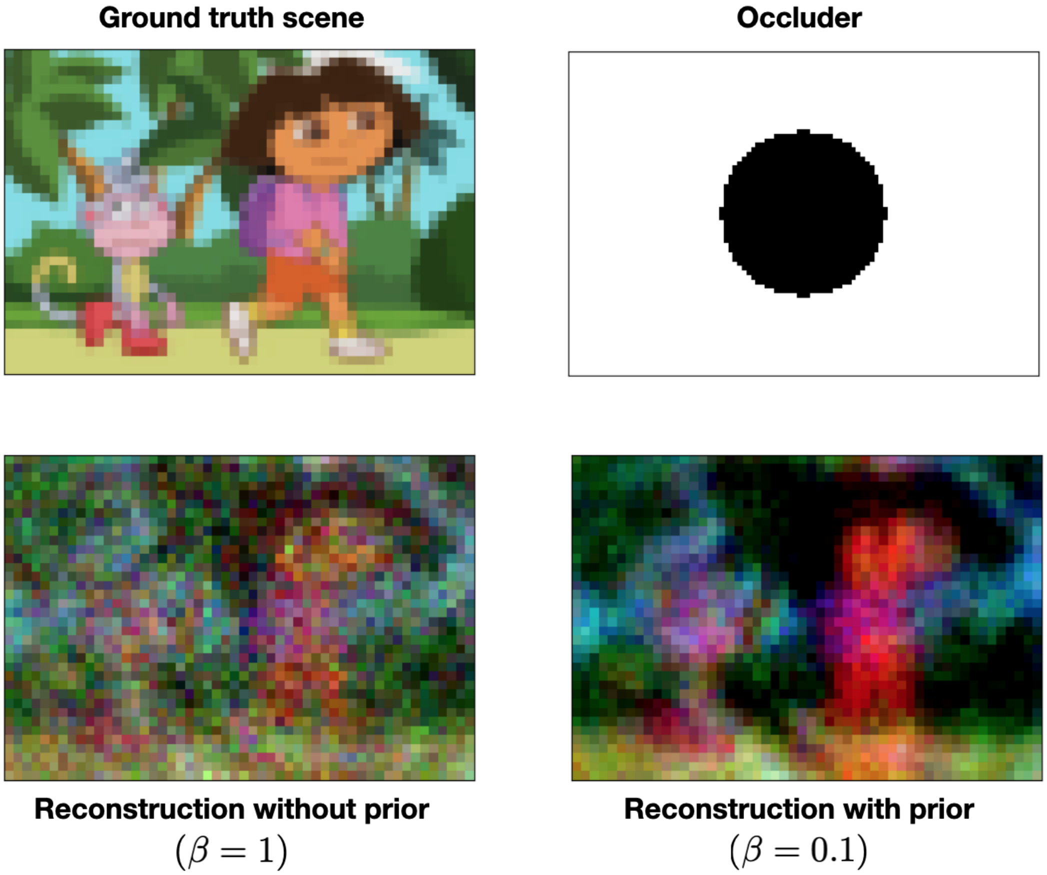


Figure 3-9: Reconstructions of a simulated scene with (lower-right) and without (lower-left) a prior.

This paper aims to analyze which occluders are optimal under most of the conditions described in Section 3.2; in particular, the analysis of [109] focuses on occluding frames, and it assumes parallel and planar occluder, scene, and observation, a diffuse observation plane, and the paraxial approximation, in flatland. It also restricts its analysis to "circulant occluders," that is, occluders that repeat themselves once, thereby yielding a circulant transfer matrix. The analysis of [109] is related to other work that considers the impact of coded apertures, such as [66, 82, 114, 67].

3.7.1 Results

As originally introduced in Section 1.4.3, we are considering two separate sources of noise: the first, W, being a fixed source of noise ("ambient" noise), and the second, J, scaling with the amount of light from the scene. In fact, in this paper, J is straightfowardly taken to be the mean total intensity of the scene itself, and so we get a number of photons arriving at the observation plane behaving like a Gaussian with mean and variance ρJ from the limit of a Poisson distribution. This paper also introduces θ, which is the variance of the scene intensity. Though this may seem counterintuitive, this corresponds to the "signal" in the signal-to-noise ratio; the variance in the scene is the information we are interested in, and the higher θ is relative to the sources of noise, the higher a fraction of the variation on the observation plane is due to information we care about.

The ρ used above is the *transmissivity* of the aperture frame; it's the fraction of the light from the scene that the aperture frame lets through. In the case of an occluding aperture frame (that repeats itself once), the transmissivity is simply the fraction of the occluder that lets through light. For example, an occluder characterized by $n = 7$ patches, where 4 of those patches let through light and 3 didn't, would have a transmissivity of $\rho = 4/7$.

As described previously: the mutual information (MI) between the measurements $y_j, j \in [n]$ and the unknowns $f_i, i \in [1, n]$ of the imaging problem is given as

$$\mathcal{I} = \sum_{i=1}^{n} \log\left(\frac{1}{W + \rho \cdot J} \cdot d_i \cdot \frac{|\lambda_i(\mathbf{A})|^2}{n} + 1\right), \tag{3.9}$$

where $\lambda_i(\mathbf{A})$ denotes the eigenvalue of $\mathbf{A}$ corresponding to the i^{th} frequency and d_i denotes the i^{th} entry on the diagonal of $\mathbf{D}$. We often write λ_i when clear from context.

Throughout this subsection we study the IID scene model. It is sometimes convenient to work with the *normalized* mutual information per pixel $\overline{\mathcal{I}} := \mathcal{I}/n$. This makes sense particularly for IID scenes, since scene-wide, the total amount of possible information to be communicated in the scene will go to infinity as $n \to \infty$. It is convenient to define $\gamma_\rho = \theta/(W + \rho J)$, for $0 < \rho < 1$. This is the SNR term. We use $\log(.)$ and $\ln(.)$ to denote the base-2 and base-e logarithms, respectively. We express all values for the MI in bits.

Pinhole

The MI of a pinhole is given by

$$\sum_{i=1}^{n} \log\left(\frac{1}{W + J/n} \cdot d_i \cdot \frac{1}{n} + 1\right).$$

In the case of an i.i.d. scene ($d_i = 1$) this further simplifies to the following

$$\mathcal{I}_{\text{pinhole}} = n \log\left(\theta/(n(nW + J)) + 1\right). \tag{3.10}$$

where, we use $\overline{\mathcal{I}} = \frac{1}{n}\mathcal{I}$ to denote the normalized MI per pixel. This is easily derived from the frequencies of an impulse. By allowing only a fraction of $1/n$ of the light to go through, the formula justifies that the performance of a pinhole deteriorates drastically for large n (cf., MI goes to zero, unless either W or $1/J$ becomes negligible, e.g., unless it scales inversely proportionally to $1/n$). In the coming sections, our MI analysis will further justify that the problem is solved by using patterns with higher transmissivity value ρ (in particular, constant with respect to n). Note that this result applies only to a vanishingly small pinhole (decreasing in size as n increases); see Section 3.13 for an analysis of fixed-size pinholes, which are the more physically relevant object.

Lens

Treating the lens transfer matrix as the identity (not scaling with $1/n$ as occluder-based transfer matrices do), the *normalized* MI of a lens is given by $\frac{1}{n}\overline{\mathcal{I}}_{\text{lens}} = \log(\frac{\theta}{W+J}+1)$. This is easily derived from the fact that the eigenvalues of the identity matrix are all 1. Thus lenses outperform any purely absorbing (mask-based) aperture.

Spectrally-flat patterns

In this section, we show that the spectrally-flat patterns maximize the MI criterion for IID scene models and dominant noise of Type-I. This is summarized in the following proposition.

Proposition 3.7.1. *Consider the IID scene model. Let $\mathcal{I}_{SF}$ be the mutual information of a spectrally-flat pattern for an odd n.*[2] *It holds that:*

$$\lim_{n\to\infty} \mathcal{I}_{SF} = \log\left(\frac{\gamma_{1/2}}{4}+1\right) + \frac{\gamma_{1/2}}{4\ln(2)}. \tag{3.11}$$

Note that this is a separate result from the result of Section 3.2; though both conclude that spectrally-flat patterns are optimal, this result is actually the stronger of the two results. The result of Section 3.2 considers the limit of high SNR; this result only requires that ambient noise dominate shot noise ($W \gg J$) without considering the strength of the signal.

Proof sketch. The proof follows straightforwardly from the fact that the first eigenvalue (i.e. the DC frequency) of the transfer matrix is $(n+1)/2$, and every other eigenvalue has a magnitude of $\sqrt{n+1}/2$, as explained in Section 3.2.2. Consider Equa-

[2] Here, we implicitly assume that n is such that a spectrally flat pattern is known to exist—see Section 3.2.2 for which values of n have that property.

tion (3.9). Because the scene is taken to be IID, we have $d_i = \theta/n$. Moreover, because the occluder is spectrally flat, we have $|\lambda_1(\mathbf{A})| = (n+1)/2$ and $|\lambda_k(\mathbf{A})| = \sqrt{n+1}/2$.

In the limit of $n \to \infty$, we have $(n+1)/2 \approx n/2$ and $\sqrt{n+1}/2 \approx \sqrt{n}/2$. Recall also that $\gamma_\rho = \theta/(W + \rho J)$. Hence, the first term in the sum given by Equation (3.9) is due to the DC term, and for that term, the factors of n cancel and we are left with $\log(\gamma_{1/2}/4 + 1)$.

Every other term in the sum is identically equal to $\log(\gamma_{1/2}/(4n)+1)$, and there are $n-1$ such terms. Because we are expressing the final mutual information in bits, the $\log(\cdot)$ function is assumed to be base 2. Using the Taylor series-based approximation of $\log(\epsilon + 1) \approx \epsilon/\ln(2)$ for vanishingly small ϵ, each of the terms in the sum is identically equal to $\gamma_{1/2}/(4n\ln(2))$ as $n \to \infty$, and because there are $n-1$ terms, we are left with a contribution of $\gamma_{1/2}/(4\ln(2))$ to the mutual information from every term in the sum save the first. This concludes the proof. □

3.7.2 Random on-off patterns

We study the MI performance of random patterns under different scenarios on the distribution of the scene and on the noise type. Our theoretical results in this section use tools from random matrix theory (RMT) and are thus asymptotic in nature. However, we include numerical simulations that show accuracy of the predictions for n on the order of a few hundreds.

Our reason for studying random on-off patterns is because as we'll soon see, under conditions different from those we just studied, we should expect the optimal transmissivity of the occluder to move away from 0.5, and under such conditions, spectrally-flat patterns are no longer easily found. (Not always impossible to find, as we'll see later in the thesis, but much rarer.) Random on-off patterns tend to do worse than spectrally-flat patterns under low-correlation or high-SNR conditions, admittedly, because of their different spectral properties; however, it's easy to generate them at any transmissivity $\rho = p$ we like, simply by making each occluder patch independently let through light with probability p, for any p we want. Hence, they're a useful tool for studying which occluder transmissivities are best under which conditions, by comparing random occluders with one transmissivity to random occluders with another.

This is the first result in this thesis that deals with the *expected* mutual information we'll get from *random* occluders, and not deterministic occluders like we're used to. With the tools we've used up until this point, such analysis would be infeasible as n became large, considering that the number of possible occluders to consider is exponential in n.

Fortunately, there's a key fact that's going to allow us to comfortably analyze random occluders, due to [15]. The key fact is this: circulant matrices whose entries are drawn IID from a particular family of probability distributions (that is, any distribution with mean 0, variance 1, and bounded third moment as $n \to \infty$) have eigenvalues whose magnitude are drawn IID from another known probability distribution. In particular, each eigenvalue is drawn from the probability density function $f_X(x) = |x|e^{-x^2}$ (see Figure 3-10).

Once we've established that, the proof follows from shifting and scaling the transfer matrix A as a function of p so that its entries have the property we're looking for.

IID scene

First, for IID scenes we explicitly compute the asymptotic value of the MI for random on-off and uniform patterns. The result is summarized in the proposition below.

Proposition 3.7.2 (Random on-off)**.** *Consider the IID scene model. Let* $\alpha_p = \frac{\sqrt{p(1-p)}}{2}$ *and* X *be a random variable with density* $f_X(x) = |x|e^{-x^2}$*. The mutual information* $\mathcal{I}_p$ *for a random on-off circulant system with parameter* $0 < p < 1$ *converges in probability with* n *to:*

$$\widetilde{\mathcal{I}}_p = \mathbb{E}_X \left[\log \left(\frac{p(1-p)}{W + pJ} X^2 + 1 \right) \right]. \tag{3.12}$$

Proof sketch. The proof leverages the following result of [15]. Consider a reverse circulant matrix $\frac{1}{\sqrt{n}}\mathbf{B}$ with entries $B_{ji} = b_{j+i-2 \mod n}$ and $(b_0, b_1, \ldots, b_n)$ a sequence of IID random variables with mean zero, unit variance and bounded third moment

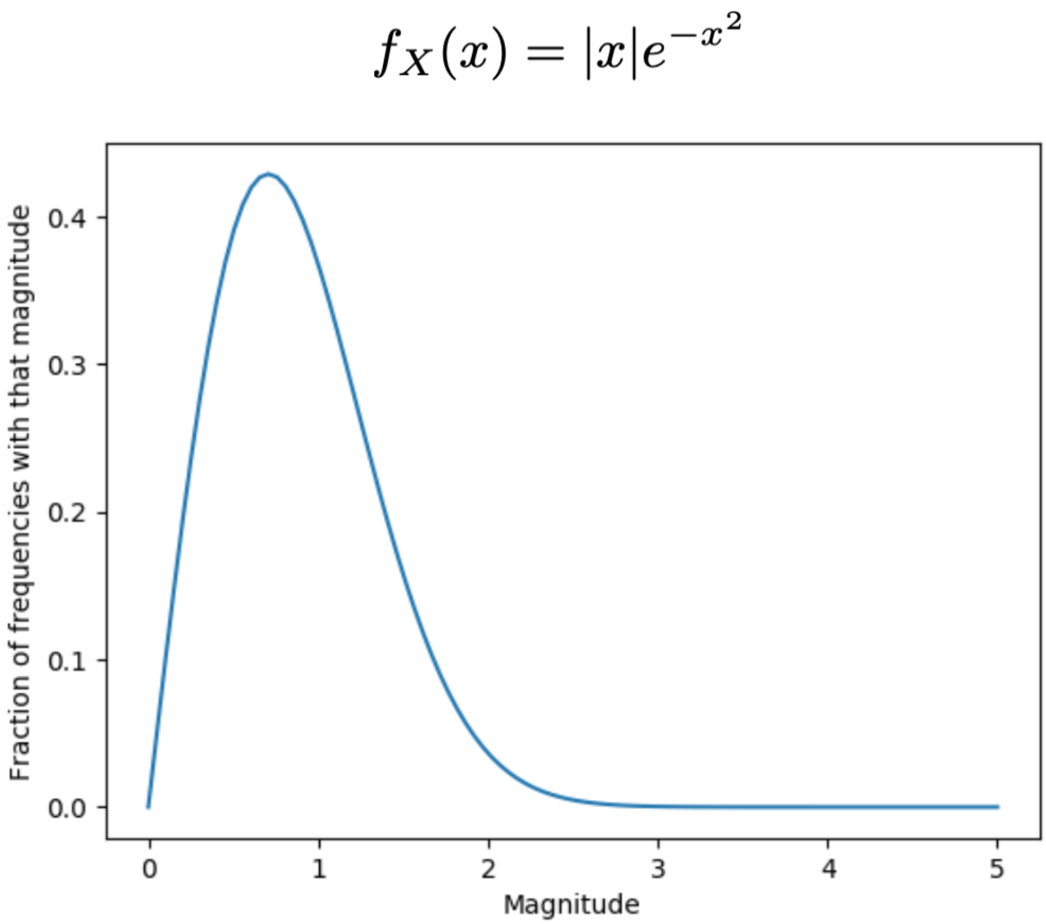


Figure 3-10: The probability density function $f_X(x) = |x|e^{-x^2}$, which characterizes the probability of getting an eigenvalue of each possible magnitude for a circulant matrix whose entries are IID with mean 0, variance 1, and bounded third moment. As $n \to \infty$, this plot will look just like a histogram of such a matrix's eigenvalue magnitudes.

(see also [14]). Then, the empirical spectral density (ESD) of $\mathbf{B}$ converges to the limiting spectral distribution with density $f_X(x)$. In our setting, we are interested on the ESD of $\mathbf{A}\mathbf{A}^T$ for $\mathbf{A}$ that has entries Bern(p). We consider the following centering of $\mathbf{A}$:

$$\widetilde{\mathbf{A}} = \frac{1}{\sqrt{p(1-p)}}\mathbf{A} - \frac{p}{\sqrt{p(1-p)}}\mathbf{1}\mathbf{1}^T. \tag{3.13}$$

The entries of $\widetilde{\mathbf{A}}$ now have zero mean and unit variance. Moreover, we have that

$\lambda_j(\widetilde{\mathbf{A}}) = \lambda_j(\mathbf{A})/\sqrt{p(1-p)}$ for $j = 2, \ldots, n$. It can be shown (see for example [15, Lem. 1]) that $|\lambda_j(\widetilde{\mathbf{A}})|^2 = \lambda_j^2(\mathbf{B})$. Applying these to (3.9) gives

$$\mathcal{I} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(1-p)}{W+pJ} \cdot \lambda_i^2\left(\frac{1}{\sqrt{n}}\mathbf{B}\right) + 1\right) \overset{n\to\infty}{\to} \widetilde{\mathcal{I}}_p,$$

where the convergence result follows from [15], as mentioned. Combining these concludes the proof of the lemma. □

Remark 1 (Optimal p). Maximizing (3.12) over p informs us on the optimal choice of the transmissivity parameter for random occluders. It can be shown that the maximum occurs at

$$p_\star = \begin{cases} \frac{1}{2}, & \text{if } J = 0, \\ 1, & \text{if } W = 0, \\ \frac{W}{J}\left(\sqrt{1+J/W} - 1\right), & \text{else.} \end{cases} \tag{3.14}$$

$$p_\star = \frac{W}{J}\left(\sqrt{1+J/W} - 1\right). \tag{3.15}$$

In particular, when ambient noise is dominant ($W \gg J$), then using $\sqrt{1+\frac{J}{W}} \approx 1 + \frac{J}{2W}$, gives $p_\star \approx \frac{1}{2}$. On the other hand, when shot noise is dominant ($J \gg W$), then $p_\star \approx \sqrt{\frac{1}{J}}$; thus, fewer open holes in the aperture design are desirable. This makes sense; if the dominant source of noise is coming from the scene, then letting in less light from the scene is better. Moreover, if the dominant source of noise is not coming from elsewhere, and the SNR is very low, then letting in lots of light from the scene is good, because it gives a better estimate of the total intensity of the scene.

Remark 2 (Comparison with spectrally-flat designs). Using Lemma 3.7.2 we can compare the MI performance of random Bern(1/2) masks with that of spectrally-flat designs, which is computed in Lemma 3.7.1. This is illustrated in Figure 3-11.

Remark 3. In this section, we saw that for IID scenes, when ambient noise is dominant,

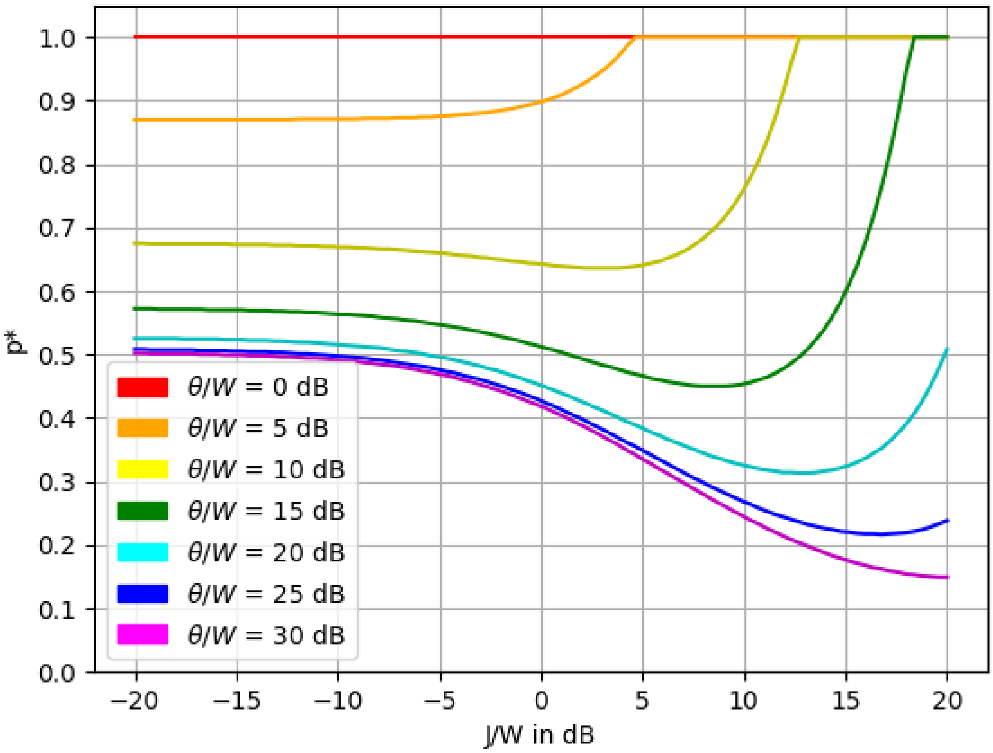


Figure 3-11: A plot of optimal transmissivity parameter $p^\star$ for random on-off patterns in the IID scene model, as we vary shot noise power J and signal strength θ while ambient noise power W is held fixed. See Remark 1.

a flat spectrum is optimal. However, this doesn't always hold true for correlated scenes. In the next section, we study the performance of different types of random masks for correlated scenes.

3.7.3 Random uniform patterns

We evaluate the MI performance of random uniform patterns as presented in the proposition below. Similarly to Proposition 3.7.2, we leverage results of [15] to evaluate the MI performance of random uniform patterns; we omit the details due to space limitations.

Proposition 3.7.3. *Consider the IID scene model. The mutual information* $\mathcal{I}_{uniform}$

for a random uniform circulant system converges in probability with n to:

$$\widetilde{\mathcal{I}}_{uniform} = \mathbb{E}_X \left[\log\left(\frac{1/24}{W+J/2}X^2+1\right)\right]. \tag{3.16}$$

$$\widetilde{\mathcal{I}}_{uniform} = \log\left(\frac{\gamma}{4}+1\right)+\frac{\gamma}{12\ln(2)}, \tag{3.17}$$

$$\gamma = \frac{\theta}{W+J/2}. \tag{3.18}$$

The proof of this proposition proceeds identically to the proof of Proposition 3.7.2, taking each element of the system to be a random uniform variable on $[0,1]$ instead of a Bernoulli random variable, and accounting for the fact that its variance is $1/12$ in the former case (as opposed to $p(1-p)$ in the latter case).

Comparing the formula of the proposition to Proposition 3.7.2, reveals that

$$\widetilde{\mathcal{I}}_{\text{uniform}} < \widetilde{\mathcal{I}}_p, \quad \text{for all} \quad p \in \left[\frac{1}{2}, \frac{1}{2}+\frac{1}{\sqrt{6}}\right]. \tag{3.19}$$

In particular, at high-SNR ($W + J/2 \ll 1$) it holds $\widetilde{\mathcal{I}}_{\frac{1}{2}} - \widetilde{\mathcal{I}}_{\text{uniform}} \approx \log(6)$. Hence, random on-off masks in this range of p outperform random uniform masks. In short, if physical limitations prevent the use of apertures that can redirect light, but can only absorb it, then absorbing all or nothing (with appropriate p) is better than partial absorption (at least for random designs). This is relatively unsurprising—while other publications [7, 20] have found occasional minor improvements to be possible under certain conditions by choosing a single element of an occluder array to be in between 0 and 1, in general we should expect occluders with mostly 1s and 0s to outperform occluders with many elements in between—after all, we want different shadows cast by the occluder to look as different as possible, and when many elements are between 0 and 1, we are throwing away some of that potential difference.

I (Adam Yedidia, not necessarily my co-authors on [109]) conjecture that as $n \to \infty$, there will exist a maximal-determinant $n \times n$ circulant matrix with $O(n)$ entries between 0 and 1 exclusive (so $O(1)$ such entries in a single row), over all possible $n \times n$ circulant matrices with entries between 0 and 1 inclusive.

In other words, when trying to solve the maximal determinants of binary circulant matrices, it won't help to set more than $O(1)$ of the entries in a row to be between 0 and 1, as $n \to \infty$.

3.7.4 Correlated scene

We extend the "worst-case" analysis of the previous section regarding IID scenes to correlated ones. We follow the β^{-d} scene prior model. We restrict the exposition to spectrally-flat and random on-off patterns. For convenience, we assume n is odd, though this will become unimportant in the limit of large n.

Introducing a non-identity covariance matrix over scenes makes the model more realistic, and makes the notion of limiting behavior in n coherent, but introduces substantial difficulties analytically. For this reason, we won't always be able to prove the results we want as cleanly as we could in the previous section; we'll occasionally have to rely on unproven conjectures.

Spectrally-flat patterns: For large enough n, we find that the MI of spectrally-flat patterns corresponds to:

$$\lim_{n\to\infty} \mathcal{I}_{\text{SF}} = \log\left(\frac{\gamma_{1/2}}{4} + 1\right) + \frac{\gamma_{1/2}(1-\beta)}{4\ln(2)\ln(1/\beta)}. \tag{3.20}$$

The derivation of this result is a straightforward extension of (3.11).

Remark 4. In the $\beta \to 1$ limit, correlations between the x_i approach 0, and it can be shown that the formula above approaches that in (3.11), as expected.

Random on-off patterns

Contrary to the case of IID scenes where knowledge of the limiting spectral density of $\mathbf{A}$ suffices to characterize the MI, for correlated scenes each eigenvalue is weighted differently. Hence, the behavior of the MI depends on the statistics of each individual eigenvalue. Since $\mathbf{A}$ is circulant, the eigenvalues of A are exactly the Fourier coefficients of the entries of the generating vector $\mathbf{a}$, i.e., $\lambda_1 = \sum_{\ell=0}^{n-1} a_\ell$, and, for $k = 2, \ldots, (n-1)/2$ (assume n is odd for simplicity): $\lambda_k^2 = \lambda_{n-k}^2 = g_k^2 + h_k^2$, where

Specifically, (assume n is odd for simplicity)

$$g_k := \sum_{\ell=0}^{n-1} a_\ell \cdot \cos\left(\ell k \frac{2\pi}{n}\right), \quad h_k := \sum_{\ell=0}^{n-1} a_\ell \cdot \sin\left(\ell k \frac{2\pi}{n}\right). \tag{3.21}$$

Note the similar form of Equations 3.21 to the simpler scenario we considered earlier of the circulant pinhole, in Section 3.13.

If the a_i's were standard Gaussians, then the following statements would hold: (a) λ_1 is distributed $\mathcal{N}(0,n)$. (b) g_k's and h_k's are IID $\mathcal{N}(0,1/2)$; therefore, $\lambda_k^2 \overset{\text{iid}}{\sim} \chi_2^2/2$ where χ_2^2 denotes a chi-squared random variable with two degrees of freedom. This leads to the following conclusion:

Lemma 3.7.1. *Let the first row of a circulant* $\mathbf{A}$ *have entries drawn IID from standard Gaussians and let the MI be given as in* (3.9) *for* $d_i = d_i^\star$ *and for some average transmissivity* ρ*. Then,* $\mathbb{E}[\mathcal{I}]$ *equals*

$$\mathbb{E}_{G\sim\mathcal{N}(0,1)} \log\left(\gamma_\rho G^2 + 1\right) + 2\sum_{k=2}^{\frac{n+1}{2}} \mathbb{E}_{X\sim\chi_2^2} \log\left(\gamma_\rho \frac{X\beta^{\frac{k-1}{(n-1)/2}}}{2n} + 1\right). \tag{3.22}$$

For $k = 2, 3, \ldots$ *let* $\phi_k : \mathbb{R}_+ \to \mathbb{R}_+$ *be such that*

$$\mu := \lim_{n\to\infty} \sum_{k=2}^{n} \mathbb{E}_{X\sim\chi_2^2}[\phi_k(X^2/2)] < \infty,$$

exists. Then, the following convergence holds:

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{k=2}^{n} \phi_k(\lambda_k^2(\mathbf{A}))\right] \overset{n\to\infty}{\longrightarrow} = \mu.$$

Define the $\mathcal{I}$ *of* $\mathbf{A}$ *as in formula* (3.9)*. Then,*

$$\mathcal{I} \overset{n\to\infty}{\longrightarrow} \lim_{n\to\infty} \frac{2}{n} \sum_{i=2}^{\frac{n}{2}-1} \log\left(\frac{1}{i}\frac{p(1-p)}{2}(\chi_2)_i^2 + 1\right)$$

We conjecture that in the $n \to \infty$ limit, the conclusion of Lemma 3.7.1 is universal over the distribution of the entries of $\mathbf{a}^T$, i.e., it holds for entries that have zero mean, unit variance, and bounded third moment. Based on this assumption, we posit that the expected mutual information $\mathbb{E}[\mathcal{I}_p]$ for a random on-off circulant system with parameter $0 < p < 1$ for the correlated scene model is given by:

$$\mathbb{E}_{G\sim\mathcal{N}(0,1)} \log\left(\frac{1}{n}\gamma_p(\sqrt{p(1-p)}\cdot G + p\sqrt{n})^2 + 1\right) \\ + 2\sum_{k=2}^{\frac{n+1}{2}} \mathbb{E}_{X\sim\chi_2^2} \log\left(\frac{1}{2n}\gamma_p p(1-p) X \beta^{\frac{k-1}{(n-1)/2}} + 1\right). \tag{3.23}$$

Remark 5. It is apparent from inspection of (3.23) that a lower β (i.e. a more correlated scene) implies a higher $p^\star$. This effect can be observed in Figure 3-12, which also compares (3.23) against simulated data. Once again, this is to be expected; if β becomes sufficiently low, then the total intensity of the scene constitutes all of the information about the scene that there exists to be learned. Comparing Figures 3-12 and 3-11 shows us that the effects of SNR, and which source of noise is dominant, have a vastly more important effect on $p^\star$, however.

Proposition 3.7.4. *Assume the universality of Proposition 3.7.1, consider the correlated scene model and let μ_p be defined as:*

$$\mu_p := \lim_{n\to\infty} \frac{2}{n-3} \sum_{k=1}^{\frac{n-1}{2}} \mathbb{E}_{X\sim\chi_2^2}\left[\log\left(\frac{1}{W+pJ}\frac{p(1-p)}{2k^2}X^2\right)\right],$$

$$\mu_p := \lim_{n\to\infty} \sum_{k=2}^{n} \mathbb{E}_{X\sim\chi_2^2}\left[\log\left(\frac{1}{W+pJ}\frac{p(1-p)}{2k^2}X^2 + 1\right)\right],$$

$\alpha_p = \frac{\sqrt{p(1-p)}}{2}$ *and Then, the mutual information $\mathcal{I}_p$ for a random on-off circulant system with parameter $0 < p < 1$ converges in probability with n to:*

$$\widehat{\mathcal{I}}_p := \lim_{n\to\infty} \mathbb{E}[\mathcal{I}_p] = \mu_p \tag{3.24}$$

The proof of the proposition is similar to the proof of Proposition 3.7.2 and further uses Proposition 3.7.1. We numerically validate the correctness of the proposition in

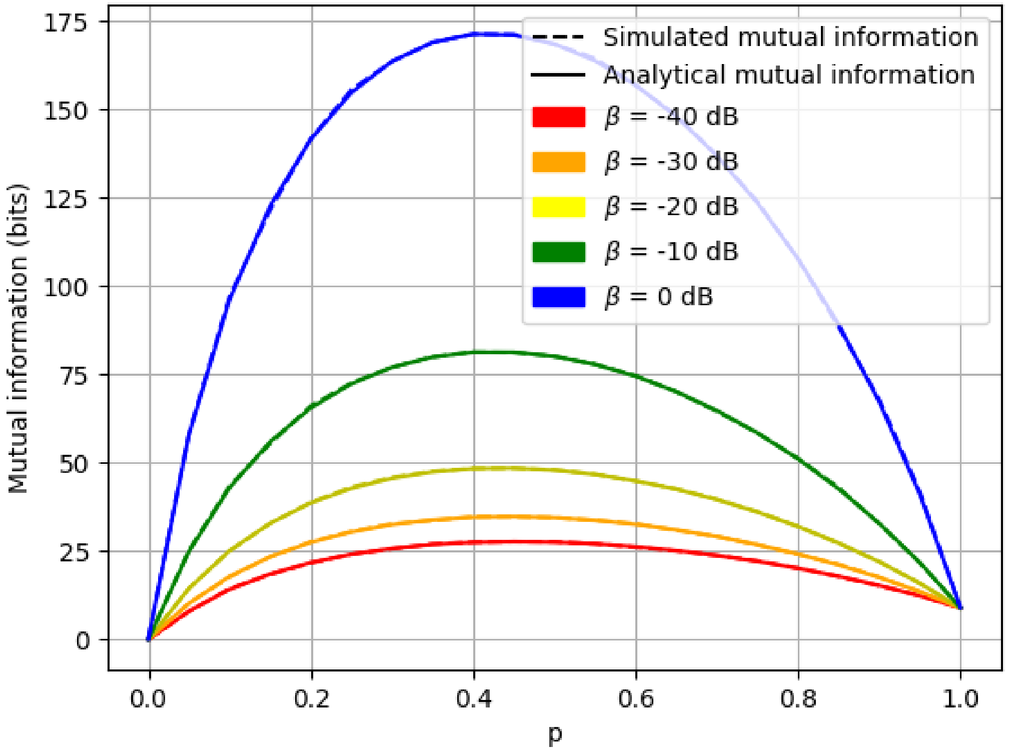


Figure 3-12: Analytical formula follows Eqn. (3.23). Simulated data are averages of 200 randomly generated apertures of size $n = 251$ for various different values p. We set $J/W = 0$ dB and $\theta/W = 30$ dB. Simulations match our analysis perfectly, providing support for the conjecture of (3.23).

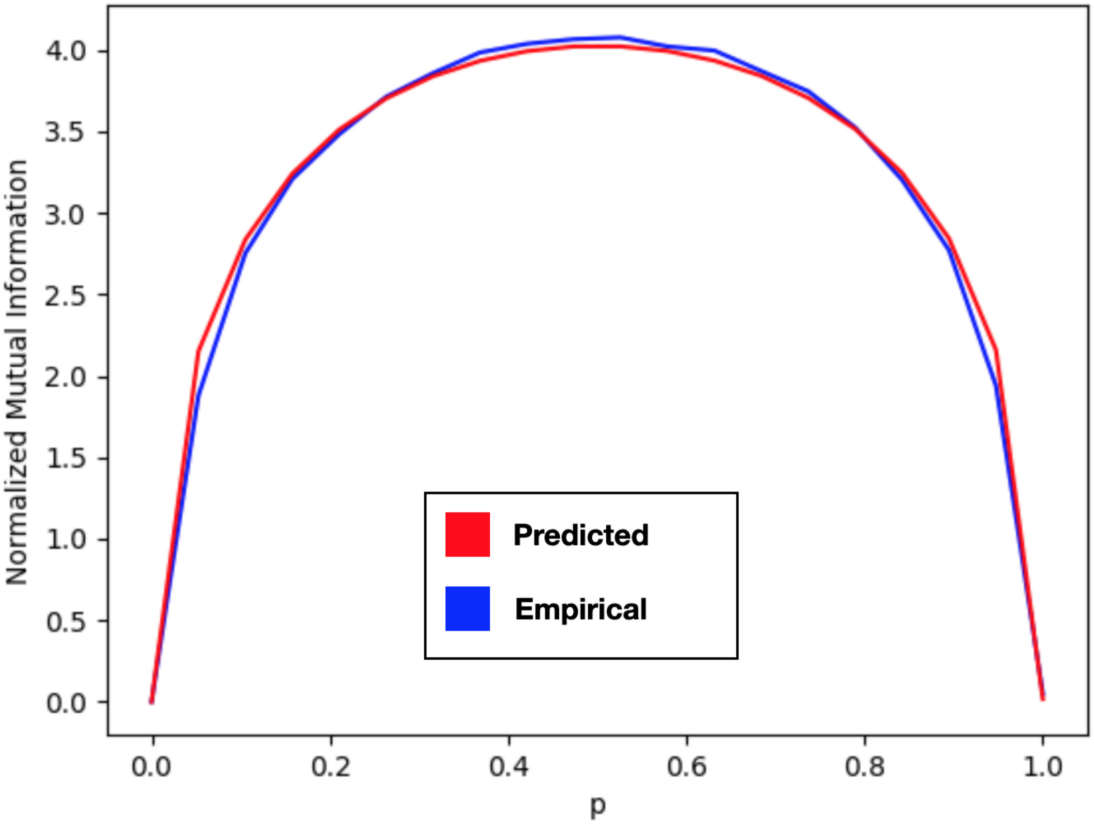


Figure 3-13: In this figure we plot the mutual information of a random on-off mask as a function of p. In blue is the simulated mutual information, obtained by generating 100 random sample masks and averaging. In red is the mutual information implied by Proposition 3.7.4. For both curves, $m = n = 1000$. The point here is we have some evidence for the truth of the conjecture we made right above Equation 3.23.

Figure 3-13.

We can use (3.24) to compute the value of p that maximizes the MI performance. Since log is increasing, this happens exactly at the value of p that maximizes the argument $p(1-p)/(W+pJ)$, i.e., for $p^\star$ as given in (3.15). Therefore, the same conclusions can be drawn as in Remark 1. Finally, using Proposition 3.7.4 we can compare the MI of random Bern(()1/2) pattern (which is optimal for $J \gg W$, cf. Remark 1) to the performance of spectrally flat designs for correlated scenes.

The MI of the latter can be computed similar to Proposition 3.7.1:

$$\log(\frac{n/4}{W+J/2}+1)+2\sum_{k=2}^{\frac{n-1}{2}}\log\left(\frac{1/4}{W+J/2}\frac{1}{k^2}+1\right), \tag{3.25}$$

for large n.

Redirection techniques

The apertures that we studied thus far physically impose a maximum-value constraint of 1 on each entry of the corresponding transfer matrix $\mathbf{A}$. Here, we study a more general family of apertures, where the only constraints that are imposed are the following: (i) the entries of $\mathbf{A}$ are nonnegative; (ii) the total power along each light ray can't be increased. Thus, the column sum of each of the columns of $\mathbf{A}$ is bounded by n. Since the matrix $\frac{1}{n}A$ is then left stochastic, we have that $|\lambda_i(\mathbf{A})| \leqslant n$. Thus, from formula (3.9) and Jensen's inequality, the MI is maximized when all eigenvalues are equal to n. This corresponds to $\mathbf{A} = n\mathbf{I}$ (or, a permutation of identity), i.e., a lens. The λ_i are the eigenvalues of A. If we are considering scenes that are IID, $d_i = 1$; if we are considering scenes with a $1/f^2$ frequency spectrum, then $d_i = \frac{1}{i}$. In this case, it matters little. Because A is left stochastic, we know that $|\lambda_i| \leqslant 1$. This is because we know that $\sum_i Ax_i = \sum_i x_i$, so if $Ax = \lambda x$, then it cannot be that scaling x by λ is increasing the sum of the entries in x, so $|\lambda| \leqslant 1$. It is therefore self-evident that the mutual information will be maximized when all eigenvalues have $|\lambda_i| = 1$. This will occur when the transfer matrix values are concentrated on the diagonal: that is, when the transfer matrix is the identity, or, in the language of optics, the aperture frame is a lens.

3.7.5 Comparing the performance of spectrally-flat occluders and others

One of the main takeaways of Section 3.7—in particular Subsection 3.7.2—is that random occluders really do perform substantially worse than spectrally-flat occluders.

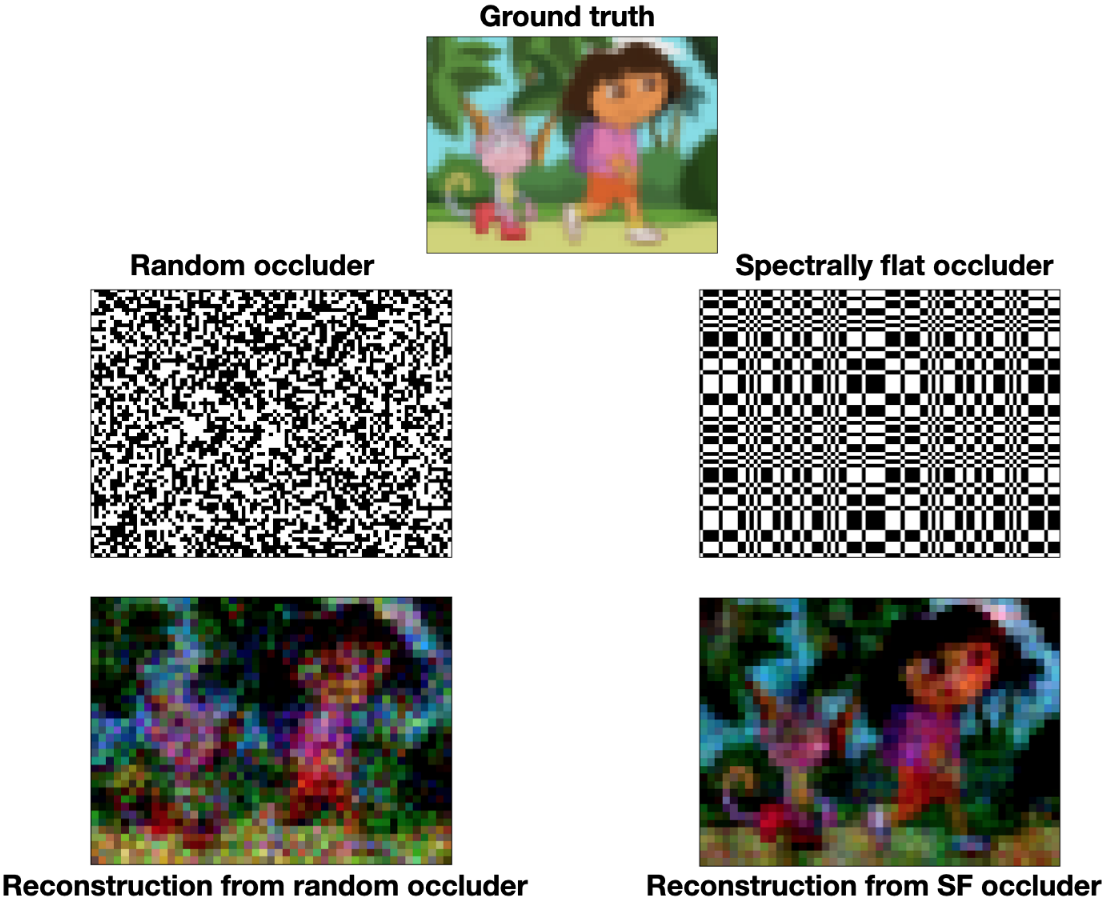

Ground truth
Random occluder
Spectrally flat occluder
Reconstruction from random occluder
Reconstruction from SF occluder

Figure 3-14: Two simulated reconstructions of a scene, using a random occluder (left) and a spectrally-flat occluder (right). The SNR is 10dB. Reconstructions are performed using the procedure described in Section 3.6, using an exponential-decay prior with $\beta = 0.1$. As can be seen, using a spectrally-flat occluder results in substantially higher reconstruction quality.

Contrary to intuition, random on/off occluders do *not* approach spectral flatness, but rather have a limiting spectral distribution of $f_X(x) = |x|e^{-x^2}$ (as explained in Section 3.7.2). To give the reader a sense of exactly how much deterioration in quality this causes in practice, see Figure 3-14.

Another salient point of comparison is with the pinhole occluder. Pinhole occluders have gained outsize prominence as occluder-based cameras for historical reasons [56] and in particular because no computation is required in order to use them; the observation itself provides an adequate reconstruction of the scene. However, they nevertheless perform importantly worse than spectrally-flat occluders.

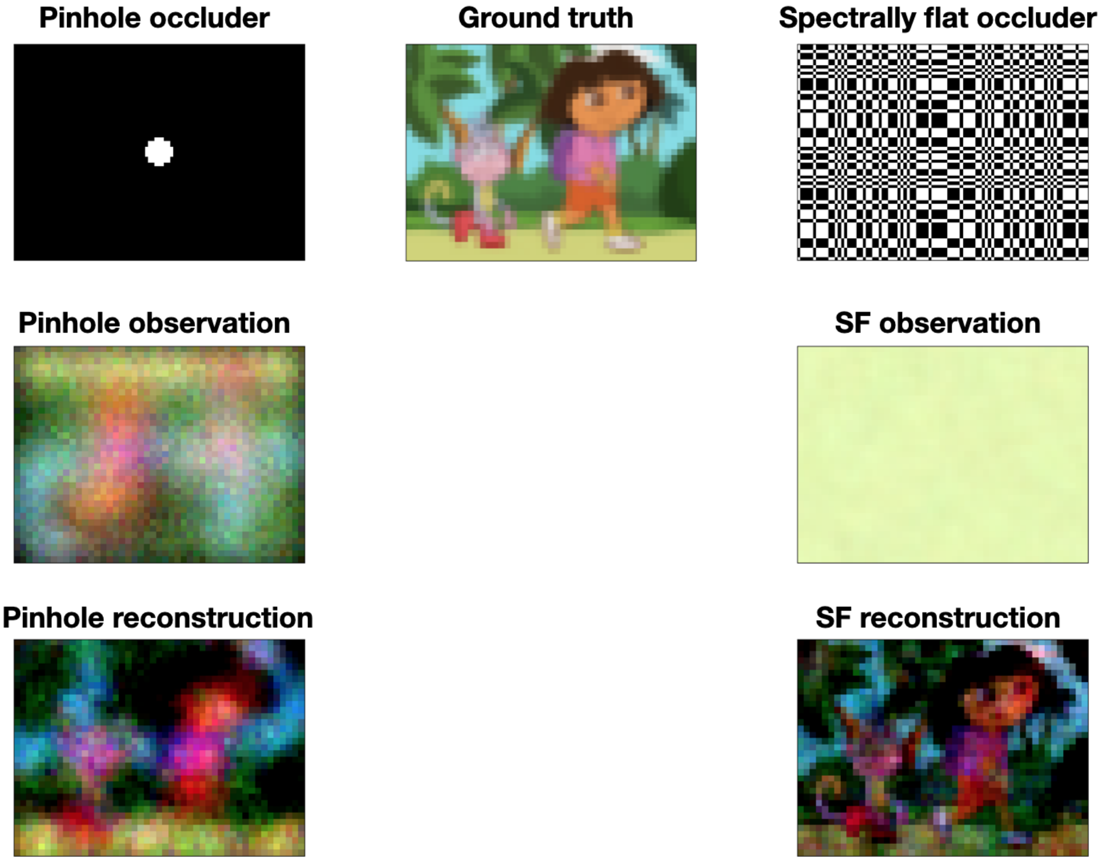


Figure 3-15: Two simulated reconstructions of a scene, using a pinhole (left) and a spectrally-flat occluder (right). The SNR is 10dB. Reconstructions are performed using the procedure described in Section 3.6, using an exponential-decay prior with $\beta = 0.1$. As can be seen, using a spectrally-flat occluder results in substantially higher reconstruction quality. Note that the observation using a pinhole is much more intelligible than that obtained using a spectrally-flat occluder, but this does not translate to higher reconstruction quality.

3.8 Near-Optimal Coded Apertures for Imaging Via Nazarov's Theorem

This section is devoted to a brief discussion of *Near-Optimal Coded Apertures for Imaging Via Nazarov's Theorem* [7]. This paper builds upon the work of [109] and contributes many useful concepts and ideas. Among them:

1. Describing spectrally-flat sequences with other transmissivities.
2. Using mean-squared error (MSE) rather than mutual information as a metric to optimize.
3. Using exposure time as an additional parameter of comparison for different imaging systems (analogously to how SNR and scene correlation are used in this thesis).
4. With mean-squared error as a metric, proving a bound on the performance of a greedy algorithm similar to the one described in Section 3.3; [7] shows that the performance of the occluders found by the greedy algorithm is within a constant factor of the performance of the optimal occluder. This constant factor is in terms of the exposure time required to get an equivalent MSE.

Of course, this extremely brief summary does not do the paper justice; I encourage the reader to read the paper itself. However, I'll go into a little bit more detail about the first two points.

3.9 Spectrally-flat sequences at different transmissivities

Section 3.2.2 describes several different constructions for binary spectrally flat sequences with ρ approaching 0.5 as $n \to \infty$ (meaning that the fraction of entries in the sequence that are 1 approaches 0.5). Section 3.7 described conditions under which the ideal occluder would have a transmissivity ρ other than 0.5; because no spectrally-flat sequence constructions are known for most values of ρ, it transitioned to analyzing random sequences as potential occluders rather than spectrally-flat sequences. However, if spectrally-flat sequence constructions were known for other values of ρ than 0.5, they would of course be preferable to random sequences.

In [7], spectrally-flat sequence constructions for $\rho = 1/4$ and $\rho = 1/8$ are given, which are loosely analogous to the Paley construction [77], but use quartic and octic residues, respectively, rather than quadratic residues. (Quartic and octic residues measure whether or not something is a power of 4 or 8, respectively, modulo n.) This is potentially useful for IID scenes where a low transmissivity is desirable; for instance, if shot noise was the dominant source of noise, as we saw in Section 3.7. These constructions are originally due to [29] and [64], respectively.

Unfortunately, unlike for the $\rho = 1/2$ constructions, it is not known whether there are an infinite number of ns for which a spectrally-flat sequences exists for $\rho = 1/4$ and $\rho = 1/8$; in the case of $\rho = 1/8$, so few are known that the practicality of this observation is diminished, with the two smallest known sequences being $n = 73$ and $n = 26041$. See Figure 3-16 for examples of spectrally-flat sequences with those transmissivities.

Note that, like in Section 3.7, the focus is on circulant occluders; the above constructions would repeat themselves once if they were to be used in an imaging system.

3.10 Discussion of two occluder scoring metrics

It's not obvious that the best occluder-scoring metric would be mutual information, and since the publication of [109], there has been further work in [7] that instead primarily relies on minimizing mean-squared error, rather than maximizing mutual information. This is a reasonable choice; the purpose of this section is to briefly discuss the impact of this choice.

Here are the two metrics, for an IID scene with signal strength θ:

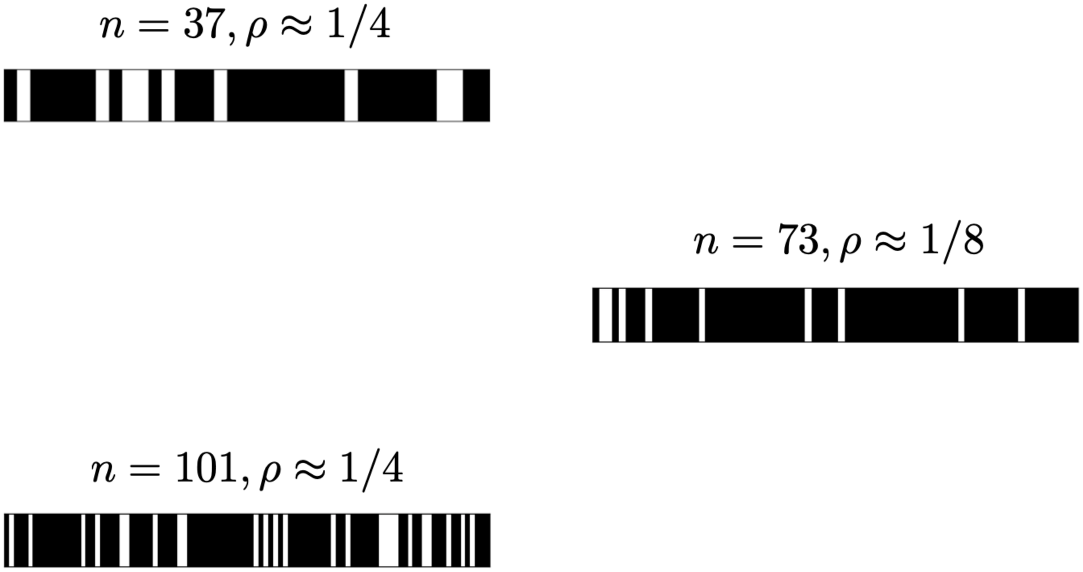

$n = 37, \rho \approx 1/4$
$n = 73, \rho \approx 1/8$
$n = 101, \rho \approx 1/4$

Figure 3-16: Three spectrally-flat sequences with transmissivities not equal to 1/2.

$$MSE_{min}(n, t, W, J, \lambda) = -\sum_{i=0}^{n-1} \frac{1}{\frac{1}{\theta} + \frac{t|\lambda_i|^2}{n(W+J\rho)}}$$

$$MI_{max}(n, W, J, \lambda) = \sum_{i=0}^{n-1} \log\left(\frac{\theta}{nW + \rho J}\frac{|\lambda_i|^2}{n} + 1\right)$$

In the above, the λ_i denote the frequencies of the occluder (or eigenvalues of the transfer matrix). We're assuming a circulant transfer matrix, and in the second formula, we've subsumed the exposure time into the signal strength (equivalently, we can take $t = 1$ in the first formula).

The two metrics are different, of course, but they're very similar in many important ways. For one thing, they both privilege spectrally-flat occluders over all alternatives in the limit of high SNR. But even at lower SNRs, they agree closely. To give a sense of how closely, Figure 3-17 shows a pair of scatter plots. As we can see from the figure, the two metrics are very highly correlated with each other.

That doesn't tell us quite what each metric is optimizing for, though. As discussed earlier, one advantage of the MSE-minimizing metric is that it's much clearer what

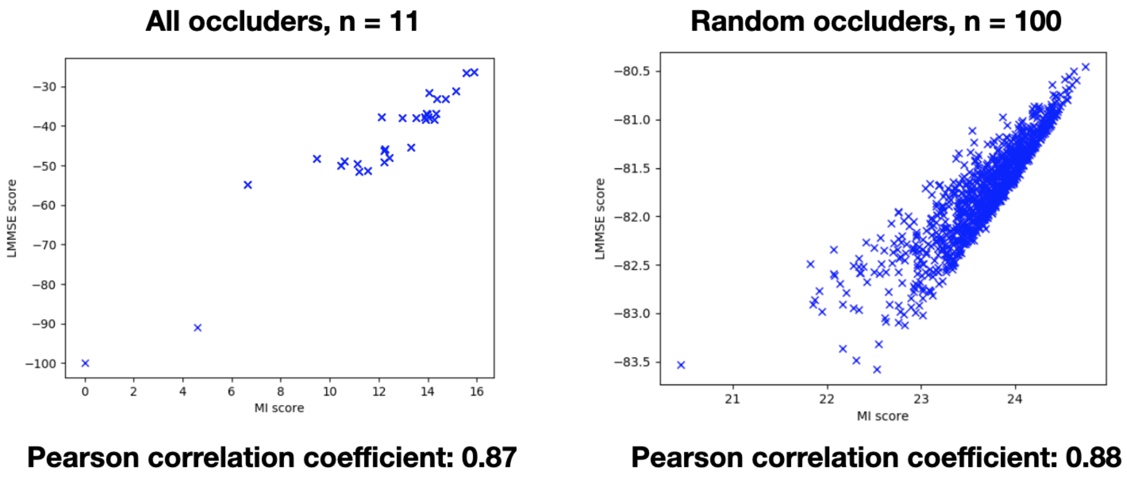


Figure 3-17: A pair of scatter plots comparing the LMMSE-minimizing metric to the MI-maximizing one. In each case, we have $t = 1, W = 1, J = 0, \theta = 100$. In the left plot, we have one point for every 11×11 binary circulant transfer matrix; in the right, we have one plot for each of 1000 randomly-chosen 100×100 binary circulant transfer matrices.

it's doing. To compare the two metrics in a more concrete way, I found a pair of occluders, one preferred by one metric, one preferred by the other. (Because the two metrics are so closely correlated, both are still about equally preferred by both!) But the subtle differences in reconstruction quality are illuminating. See Figure 3-18 for a side-by-side comparison of reconstructions of the same scene under the same conditions, but using each of these two different occluders.

Evaluating reconstructions such as these, beyond simply reporting their mean-squared error, is a subjective business. As we expect to see, the reconstruction using the occluder that is better for reducing the reconstruction MSE has the lower MSE. But the other reconstruction does seem a bit sharper it is instructive to see what it might mean to increase the "information content" of the reconstruction, and reassuring that it seems to correspond to something meaningful. Moreover, The scatter plot of Figure 3-17 suggests that the choice between maximizing mutual information or minimizing MSE matters relatively little.

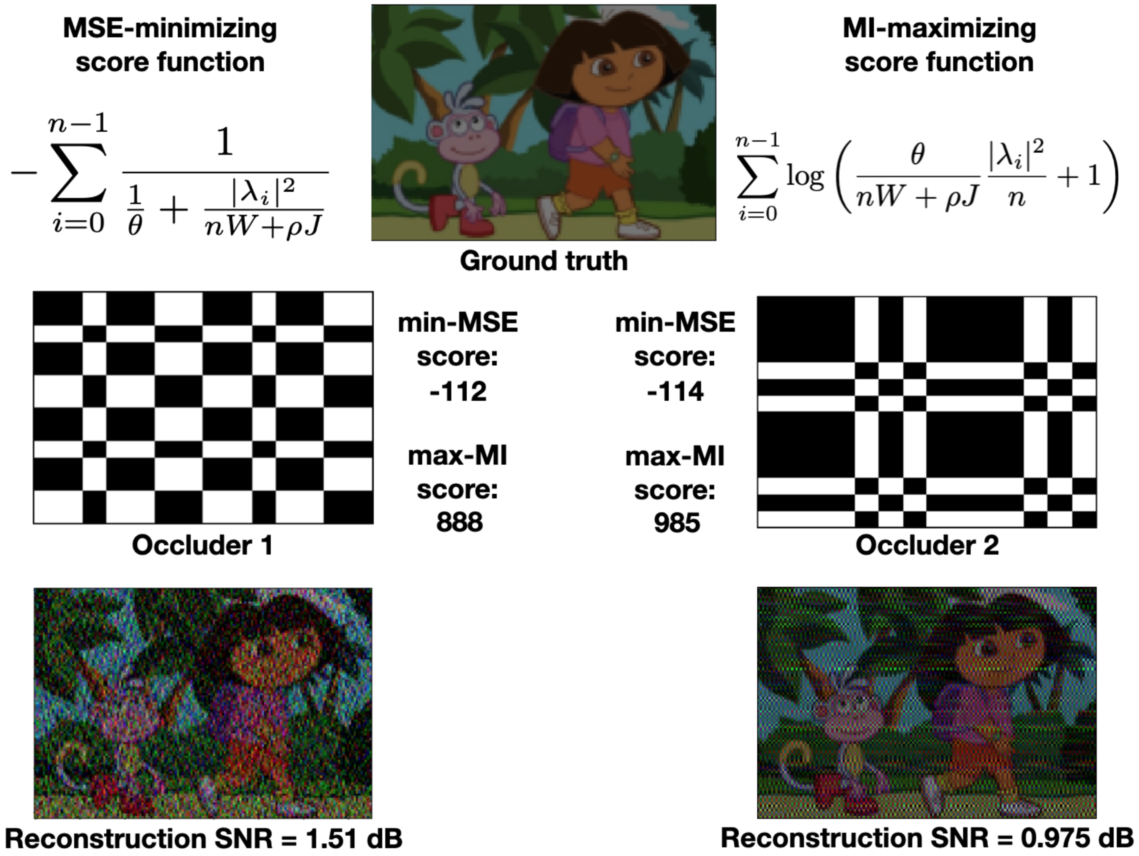

MSE-minimizing score function
$-\sum_{i=0}^{n-1} \frac{1}{\frac{1}{\theta} + \frac{|\lambda_i|^2}{nW+\rho J}}$
Ground truth
MI-maximizing score function
$\sum_{i=0}^{n-1} \log\left(\frac{\theta}{nW+\rho J}\frac{|\lambda_i|^2}{n}+1\right)$
min-MSE score: -112
max-MI score: 888
min-MSE score: -114
max-MI score: 985
Occluder 1
Occluder 2
Reconstruction SNR = 1.51 dB
Reconstruction SNR = 0.975 dB

Figure 3-18: A side-by-side comparison of reconstructions of the same scene under the same conditions, but using two different occluders: one (left) preferred by the MSE-minimizing metric, the other (right) by the MI-maximizing metric. As expected, the MSE of the right reconstruction is worse, but the image appears to be somewhat sharper.

3.11 Varying the distance between observation, occluder, and scene

Let's continue examining each of the design constraints of the idealized model. Consider the assumption that the occluder lies exactly halfway in between the scene and observation plane. This was a tremenously convenient assumption because it allowed us to assume that the occluder's transfer matrix had Toeplitz structure. But in the real world, the assumption is completely unrealistic. In a designer-mask camera application, the occluder will presumably be much closer to the camera's photosensitive material than to the scene, and even in an accidental-camera application, we can't assume that the occluder will be exactly halfway between the wall we're looking at and whatever it is we're trying to image. So let's try removing the assumption and seeing what happens. Note that we'll still be assuming that scene and occluder are both planar—we'll get to that eventually, but not yet. And we're still using the paraxial approximation—meaning that regardless of what we are taking as the *relative* distances of the occluder to the scene and observation, we are always assuming that the distance between scene and observation to be much bigger than the size of the scene or observation.

Note that we are continuing to assume that the scene and observation are the same size. If we make the scene bigger in proportion to its distance from the occluder (that is, if the ratio of the distance from the scene to the occluder to the distance from the occluder to the observation equals the ratio of the size of the scene to the size of the observation), then the tools of analysis described in this section are actually not necessary. It would suffice in that case to simply discretize the scene more coarsely in proportion to its size, in which case all of the conclusions of previous sections (and the diagram in Figure 1-10) would still apply.

What exactly is it about an occluder halfway between the scene and observation that gives us Toeplitz transfer matrices? The answer is that when the occluder is halfway bewteen the scene and observation, the shadow cast by a moving light source will move at exactly the speed the light source is moving, but in the opposite direction. Try holding a flashlight (such as one from a smartphone) with your right hand, illuminating a table or a wall, and then hold your left hand halfway in between the flashlight and the table. (I encourage you to actually do this!) Keep your left hand steady, and then move the flashlight around. You can see that your hand's shadow moves at the same speed your flashlight does, and in the opposite direction.

Now try varying the height of your left hand relative to the table. What happens to the speed of your hand's shadow relative to the speed at which you move the flashlight? The answer is that when your hand is closer to the table than to the flashlight, your hand's shadow will move more slowly than the flashlight; and when

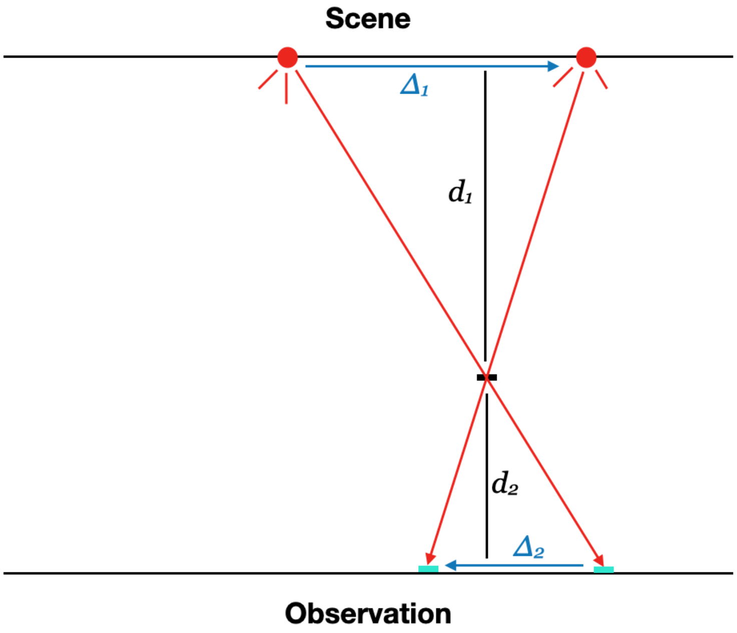


Figure 3-19: A simple layout that explains the phenomenon whereby the relative speeds of a light source and its shadow are given by the relative distances of the scene and the observation to the occluder. Suppose we have a pinspeck occluder, and a light source that moves by an amount Δ_1 to the right. If we suppose that its shadow moves by an amount Δ_2 to the left, and that the occluder has a perpendicular distance d_1 from the scene and d_2 from the observation, then the fact that the top and bottom triangles are similar tells us that $\Delta_1/\Delta_2 = d_1/d_2$.

your hand is closer to the flashlight than to the table, your hand's shadow will move faster than the flashlight. (Of course, your hand's shadow will always in the opposite direction from the shadow—that part won't change.)

In fact, to be more precise, the "speed multiplier" that your hand's shadow gets relative to the flashlight—that is, your hand's shadow's speed divided by the flashlight's speed—is the same as the distance between your hand and the table divided by the distance between your hand and the flashlight. Figure 3-19 gives a visual explanation of this phenomenon.

This "speed multiplier" concept is crucial to understanding how varying the occluder's depth warps the resulting transfer matrix. Remember that each column of the transfer matrix tells us what the observation will look like in response to a point light source at each different location in the scene. If we imagine, then, a point light source moving at a constant speed of 1 space-unit per time-unit across the scene, then we can imagine the transfer matrix as a movie of the observation plane while that happens, with each column of the transfer matrix being one frame of that movie.

When the occluder is halfway in between the scene and observation plane, we know exactly what that movie should look like: the shadow should move at the same speed as the point light source. That is, it should move at a speed of 1 space-unit (1 "bin," or $1/n$) per time unit (1 "frame," or column of the transfer matrix), assuming we discretize the scene and the observation plane equally finely.

It's for this reason that the occluder being halfway between the scene and observation gives us the perfect, constant diagonals that characterize a Toeplitz matrix. Compare a column of the transfer matrix to the column adjacent to it, and you should see a copy of that column, but shifted by one row.

What if we continue to imagine that the transfer matrix is a movie of the shadow cast by a point light source moving a constant speed of 1 space-unit per time-unit—but now we supposed that the occluder was twice as close to the scene as it was to the observation plane? We know from Figure 3-19 that that means that the shadow must move at a speed of 2 space-units per time-unit. Therefore, on the transfer matrix, moving one column (time-unit) to the right will cause the shadow to shift two rows (space-units) down. (Remember that we are sticking to our convention of labeling the observation plane right-to-left instead of left-to-right, as explained in Section 1.4—if we weren't, that would cause the shadow to shift two rows *up*!)

Similarly, if the occluder was twice as close to the observation plane as to the scene, the shadow would move at a speed of 0.5 space-units per time-unit. And if the occluder was right up against the observation plane, the shadow wouldn't move at all—and if the occluder was right up against the scene, there would be no shadow! Figure 3-20 shows some example transfer matrices for each of these scenarios.

If you look carefully at Figure 3-20—in particular the second and fourth configurations, in which the occluder is a quarter or three-quarters of the way to the scene—you'll see that the transfer matrix isn't perfectly binary, like some of the previous transfer matrices we've looked at. It contains a few entries that lie between

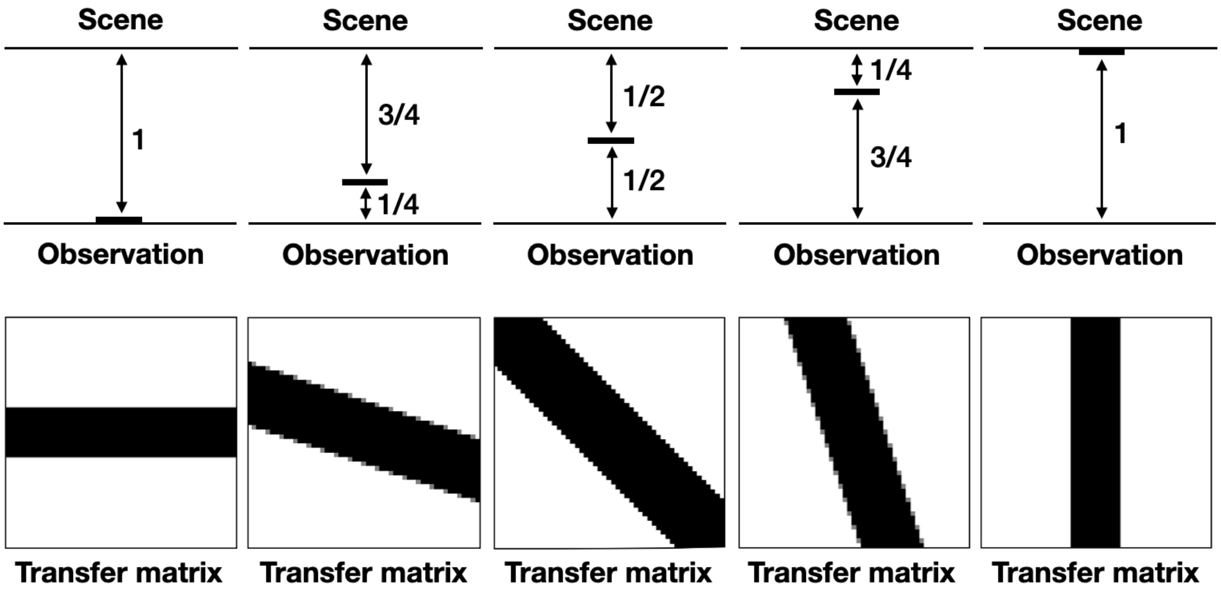


Figure 3-20: Top row: five different scenarios with the occluder at five different depths. Bottom row: the transfer matrices corresponding to each different scenario.

0 and 1. This isn't for any legitimate reason, like a partially opaque occluder; this is purely a modeling issue. It has to do with the fact that, in order to approximate the scene and observation, we've partitioned them both into discrete patches. If they were both perfectly continuous, their transfer matrices would both be completely binary, as we'd hope to see. The problem is, though, this isn't quite an issue we can wish away by appealing to what happens in the limit as our discretization becomes finer and finer: even if our discretization was extremely fine, the absolute number of nonbinary elements in our transfer matrix wouldn't shrink; in fact, it would grow! Granted, the *fraction* of nonbinary elements in our transfer matrix would shrink, but even that isn't true if we also suppose that our occluders become increasingly complex (with more and more interfaces between occluding and not-occluding). So our model remains annoyingly unfaithful to reality even when we discretize very finely.

Why is having nonbinary elements in our transfer matrix a problem? Beyond the simple fact that it doesn't accurately describe reality, it will result in our underestimating the mutual information of configurations where the occluder is not exactly halfway in between scene and occluder. This is because nonbinary elements in the transfer matrix (or the occluder, for that reason!) tend to lead to low mutual information, for reasons described in more detail in Section 3.7.3. This effect gets exacerbated when the occluder is just a little bit off from being halfway in between the scene and occluder. For example, suppose the occluder is 5/11 of the way between the observation and the scene; the resulting transfer matrix will have its diagonals of constancy be terribly skew to the diagonals of the matrix, resulting in a lot of nonbinary elements.

Fortunately, there's an easy solution to this modeling issue, and it doesn't require us to treat everything as fully continuous.[3] The key fact here is that we can relate the eigenvalues of the Gram matrix $AQA^T + I$ derived from the true, continuous transfer matrix A (which is square) to the eigenvalues of the equivalent Gram matrix $A_rQA^T + I$ derived from a rectangular version A_r of the transfer matrix A.

The way we obtain A_r from A is simply to "stretch" the matrix A (whose lines of constancy lie skew to the diagonals of the transfer matrix) until we get a version whose diagonals align with the diagonals of the transfer matrix. Figure 3-21 shows how this works.

To figure out how much to rescale in response to the stretching of the matrix, it is instructive to consider the all-ones transfer matrix (corresponding to no occlusion in the paraxial limit). This is a helpful transfer matrix for gaining intuitions about these configurations in general; it will be very helpful here.

Because the all-ones transfer matrix corresponds to no occlusion, we can equivalently posit the occluder "plane" to be at any depth. Let's imagine it's $w/(w+h)$ of the way to the scene (meaning that if the distance between scene and observation is d, then the distance between the occluder and the scene is $hd/(w+h)$ and the distance between the occluder and the observation is $wd/(w+h)$.) As Figure 3-21 shows us, that tells us that if the true transfer matrix has a width-to-height ratio of 1:1, the stretched transfer matrix will have a width-to-height ratio of w:h.

[3] Note that this is by no means the only solution to the problem of modeling continuous operators discretely [74]. Sampling theory gives us ways of approximating integral operators other than naive linear interpolation, as depicted in Figure 3-21. However, this trick of considering a rectangular proxy for a square transfer matrix is simple, and what I use in my later analysis and simulations.

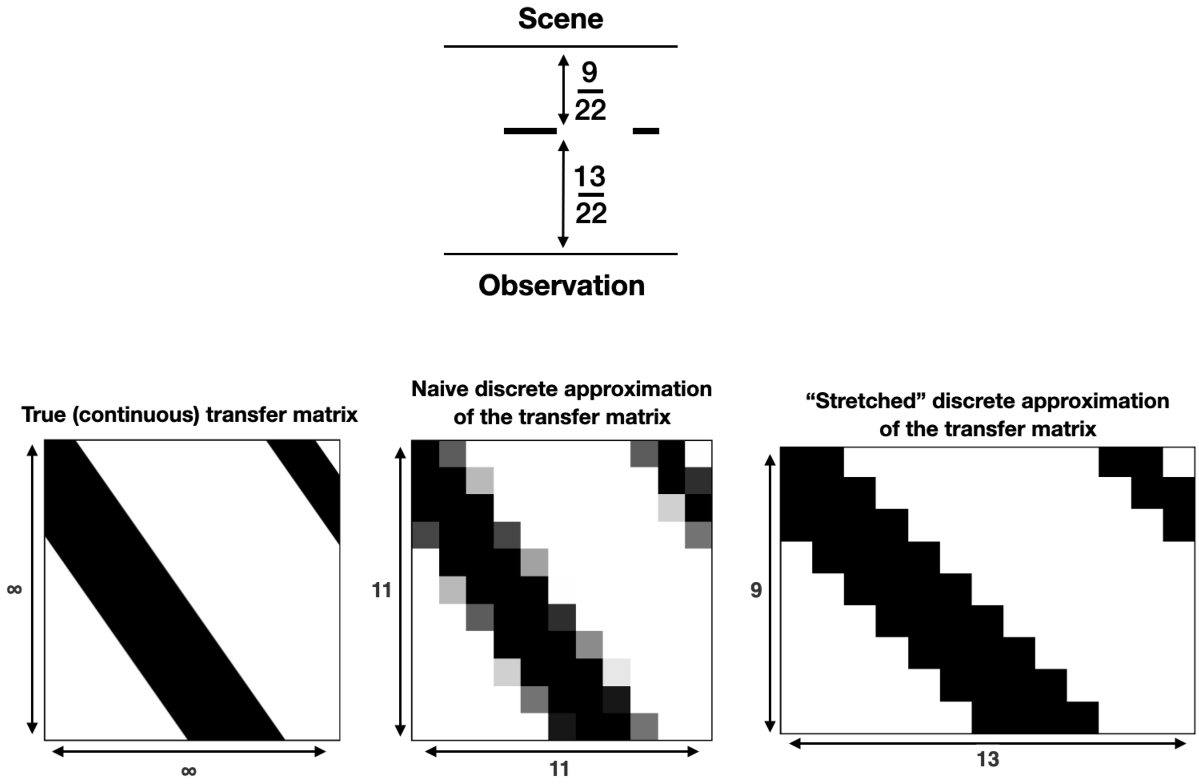

Scene
9/22
13/22
Observation
True (continuous) transfer matrix
∞
∞
Naive discrete approximation of the transfer matrix
11
11
"Stretched" discrete approximation of the transfer matrix
9
13

Figure 3-21: Top: a configuration with a planar occluder not halfway in between the scene and the observation. If we take the distance between the scene and observation to be d, then the distance between the scene and occluder is $\frac{9}{22}d$ and the distance between the occluder and observation is $\frac{13}{22}d$. We take a discretization level of $n = 11$ (so for computational reasons, we are modeling the scene and observation as vectors of $n = 11$ constant entries, and the occluder as a vector of $2n-1 = 21$ constant entries). Bottom left: the true, continuous transfer matrix. Bottom middle: the naive discrete 11×11 approximation of the transfer matrix, using averaging to produce nonbinary elements. Bottom right: the "stretched" 9×13 approximation of the transfer matrix, yielding a rectangular matrix with Toeplitz structure and only binary elements.

This has two impacts on the value of the determinant of $A^TQA + I$. The first is that $A^TQA + I$ is a bigger matrix, which will cause the value of the determinant to increase. The second is that each entry of A^TQA is smaller, because A^T is narrower and A is shorter; this will cause the value of the determinant to decrease. To rectify the first effect, it is necessary to rescale A^TQA by a factor of n^2/w^2. To rectify the second, it is necessary to rescale each eigenvalue of the A^TQA by a factor of n/h (so that taking the determinant of A^TQA+I means taking the product $\prod_i 1 + \lambda_i(n/h)$), where λ_i denotes the i^{th} eigenvalue of the (rescaled) matrix A^TQA.

To be a bit more concrete: recall our original formula for computing the mutual information of an occlusion-based system. Let A be the transfer matrix of the system, scaled to be binary-valued. Now let A_r be the rectangular equivalent of that transfer matrix, stretched so that the diagonals of constancy are the actual diagonals of the matrix. Ordinarily, we would write (as we described in Section 3.1) that the mutual information $\mathcal{I}$ is given by:

$$\mathcal{I} = \log\det(\sigma A^T Q(n)A/n^2 + I) = \log\prod_i(1 + \lambda_i(\sigma A^T Q(n)A/n^2))$$

where $Q(n)$ is the covariance matrix of the scene, σ is the signal-to-noise ratio, and $\lambda_i(A^TQA)$ denotes the i^{th} eigenvalue of A^TQA.

Instead of taking A to be a square matrix with diagonals of constancy that are skew to the actual diagonals of the matrix, we stretch it so that its width to height ratio is w:h. Let's take its dimensions to be $w \times h$. Call this new, stretched matrix A_r. Now we can write:

$$\mathcal{I} = \log\prod_i(1 + \frac{n}{h}\lambda_i(\sigma A_r^T Q(w)A_r/w^2))$$

In the limit of continuous transfer matrices, these two formulations are perfectly equivalent. But when we approximate the continuous transfer matrix by using a discrete matrix, the difference between these two "equivalent" formulations can be dramatic—and, of course, it's the rectangular formulation that gets closer to the truth. See Figure 3-22 for a side-by-side comparison.

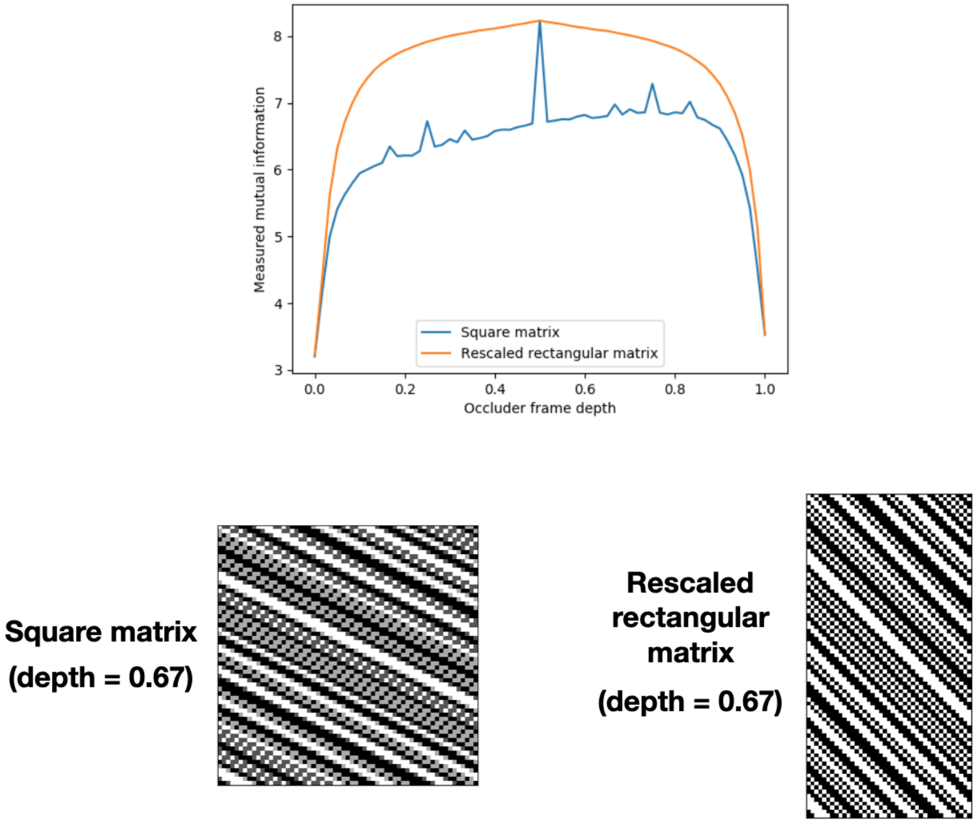


Figure 3-22: A comparison of observed mutual information using either a square matrix with non-binary averaged values, or a stretched rectangular matrix, appropriately rescaled. In this example, $n = 61$. It is apparent that the effect of averaging values in the square matrix leads to a dramatic (and artifactual) decrease in the observed mutual information, as well as artifacts at depths that make there be fewer nonbinary values in the square matrix.

This trick lets us faithfully represent systems involving occluders at $2n-1$ different depths, where n is the level of discretization of the scene. This is tremendously convenient! It lets us do an accurate study of how much having occluders not exactly halfway in between the scene and observation harms performance. We know, of course, that it must: DMBCs have diagonals of constancy in line with the matrix's diagonals, and skewing those diagonals decreases the rank of the resulting transfer matrix! But by how much is an interesting question. Figure 3-23 gives an answer. Just like Figure 3-6, it helps us estimate the effective pixel count of the configurations involving occluders at different depths. As expected, moving the occluder away from the halfway point reduces the configuration's effective pixel count—but interestingly, the effective pixel count doesn't change much until the occluder is moved far from the halfway point (e.g. 10% or 90% of the way to the scene). This tells us that using a good occluder matters a lot more than making sure it's exactly halfway between scene and observation, for the purposes of maximizing mutual information.

3.12 Near-field scenes

Next, we tackle the question of what happens to our analysis when we discard the paraxial approximation. Real scenes, after all, don't lie infinitely far away from our cameras. And though we saw in Section 1.3 that the paraxial approximation is deceptively robust, because of the quadratic dependence on distance in the illumination function (see Equation 1.1), that doesn't mean it's *right*. So let's analyze our standard configuration more carefully, without using the paraxial approximation.

Way back in Section 1.4, we decided to use a reversed labeling system, such that the scene vectors were ordered left-to-right (as normal) but the observation vectors were ordered right-to-left. This was so that the diagonals of constancy of the resulting transfer matrices would go from upper-left to lower-right. This lets us work with circulant and Toeplitz matrices, rather than Hankel matrices. Of course, in the end, it doesn't matter what labeling scheme we use: the math must work out the same in either case. However, this reverse labeling was more convenient in the previous

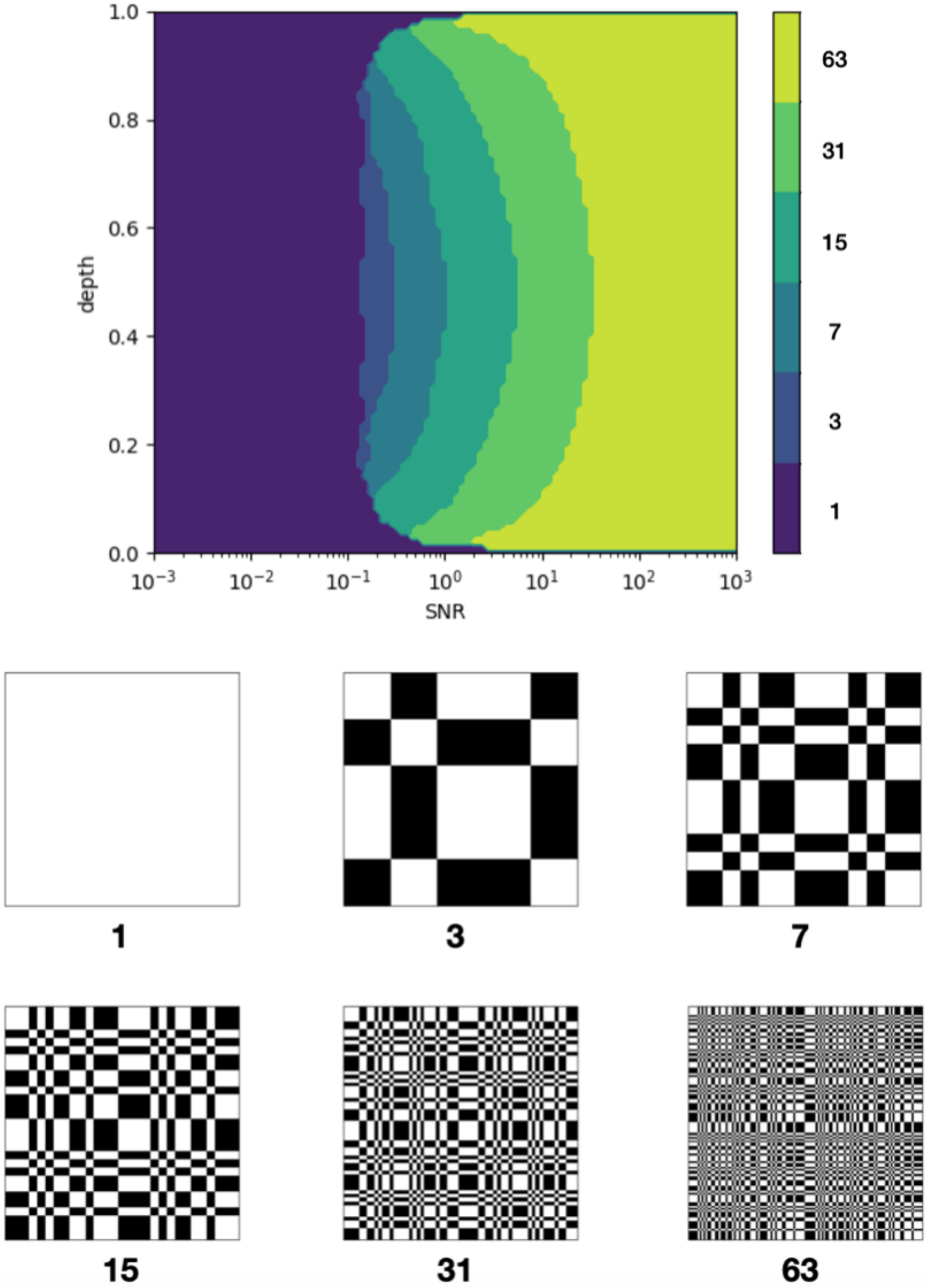


Figure 3-23: Top: the approximate effective pixel count of scenes generated at different occluder depths. As expected, when the occluder is near the observation plane or the scene, it reduces the number of effective pixels. Here the frequency attenuation coefficient $\beta = 0.1$. Higher values of β correspond to less correlated scenes. Bottom: masks corresponding to each of the effective scene pixel counts. Note that these masks repeat themselves once in each dimension, so each mask is $2n-1 \times 2n-1$ if the effective pixel count is n. This is due to the phenomenon described in Figure 3-23.

sections.

In this section, to avoid confusion, we'll stick with the same convention as we used in previous sections. Unfortunately, though, in this section, it won't buy us any convenience at all. The reason is this: the near-field effects map in exactly the opposite direction to the occlusion effects! Consider a configuration with the scene plane at $y = 1$, the observation plane at $y = -1$, and the occluder frame at $y = 0$. Suppose that the occluder frame includes an aperture (meaning no occlusion) at the point $(x, y) = (0, 0)$. Now each point on the observation is guaranteed a contribution from the point across from it in the scene; that is, a point at $(x, -1)$ on the observation is guaranteed a contribution from point $(-x, 1)$ in the scene. It's that negation that drove us to swap the index order of scene and observation, so that scene went from $-x_{max}$ to x_{max} and observation from x_{max} to $-x_{max}$.

But now let's think about the near-field effects. Ignoring the effects of occlusion, a point at $(x, -1)$ on the observation will get the most light from the point nearest to it in the scene—that is, $(x, 1)$. That light will fall off as $I = y/(x^2 + y^2)$, as we saw earlier. This is exactly the reverse of the occlusion phenomenon. So if the diagonals of constancy of transfer matrices in a world with occlusion but no near-field effects go from upper-left to lower-right, the diagonals of constancy of transfer matrices in a world *without* occlusion, but *with* near-field effects, must go from upper-right to lower-left. We just can't win!

So what happens when we have both occlusion and near-field effects? The answer is simple: we take the Hadamard product—i.e. the elementwise product—of the two transfer matrices with each individual effect. Unfortunately, not much is known that can be said about the eigenvalues of the Hadamard product of two matrices analytically. We can still say interesting things about it through a combination of common sense and simulations, however.

Let's start with the common sense. When the scene is close enough to the observation plane, the near-field effects are effectively a blurry pinhole, but one that treats the scene as reversed relative to how an actual blurry pinhole would treat the scene. (See Figure 3-24.) When we take the Hadamard product of these two matrices, it's

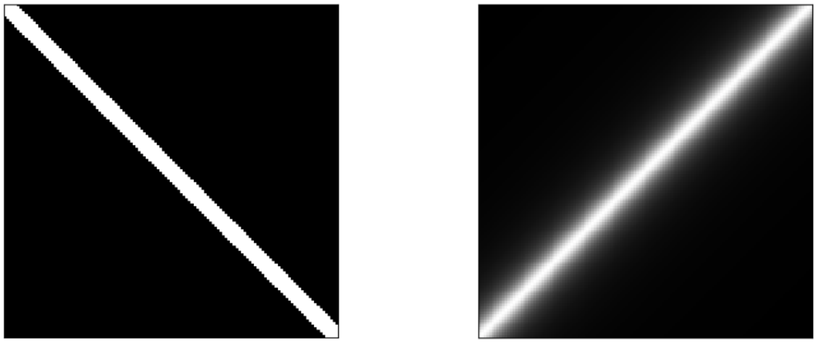


Figure 3-24: Left: the transfer matrix from a pinhole. Right: the transfer matrix from *very* strong near-field effects ($d = 0.01x$, where d and x are the distance from scene to observation and size of observation, respectively). As can be seen, the transfer matrices look identical but for a vertical (or horizontal) reflection.

intuitive that it would the determinant of the matrix that results. After all, the determinant of a matrix is a sum of permutations; when we take the Hadamard product of these two matrices whose permutations are non-overlapping, it makes sense that they would interact destructively. In other words, it won't help to have both the near-field effect and the occlusion effect happening at once; the result won't be better than either having the near-field effect alone, or the occlusion effect alone. Which one of those two is better, of course depends on the details: what occluder are we talking about, and how strong is the near-field effect?

If we're to talk about which occluder is *optimal* under these conditions, we might expect, then, that while near-field effects are weak, the optimal occluder continues to be whatever was optimal without the near-field effects; as we gradually strengthen the near-field effects, at some point the optimal occluder will suddenly switch to being a completely open aperture. This should happen once a simple pinhole starts outperforming a spectrally-flat occluder that is being marred by increasingly strong near-field effects. See Figure 3-25 for an example of a transfer matrix near-field effects strong enough to be approaching that threshold.

And indeed, that's what we see in our simulations! See Figure 3-26. As can be seen, the optimal occluder suddenly becomes the empty occluder beyond a certain threshold based on the strength of near-field effects.

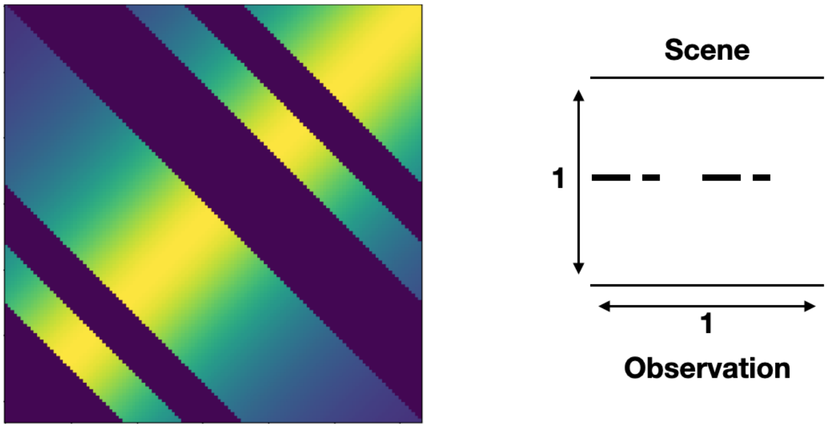


Figure 3-25: Left: the transfer matrix from an occluder-based imaging configuration with strong near-field effects. Right: the imaging configuration.

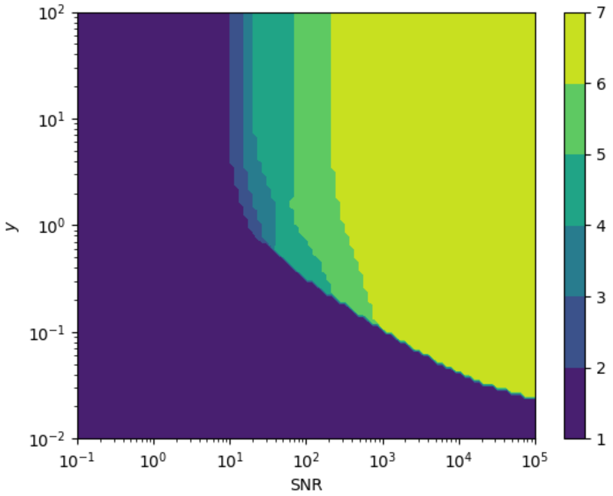


Figure 3-26: A resolution plot showing the effective pixels per side as a function of the SNR and the distance of the scene from the observation. $\beta = 0.1$. The length of $x_{\max}$ of the observation and scene is $x_{\max} = 1$.

3.13 The Optimal Pinhole

What size of pinhole is optimal, given our standard assumptions? This might seem to be an irrelevant question—the optimal occluding frame is not a pinhole, so why should we care what size of pinhole is optimal?

It is a simple enough question that we are able to solve it analytically, which is nice. It can come up that you only have a pinhole and all you can control is its size. It is a good and intuitive test case for a lot of the mutual-information-based machinery that has been introduced so far. But more importantly, it will later be useful for us to have analyzed this question, as the idea of a "wide pinhole" will be a useful analogy for near-field effects.

First of all, at a high level, what is the fundamental tradeoff around pinhole size? The answer is that smaller pinholes let in less light, but larger pinholes give you a blurrier image. If you're nearsighted, you can see this tradeoff in action by squinting—squinting lets you get sharper view of whatever it is you're looking at, but squint too much and your eyes won't get enough light to see anything! It's intuitive, then, that a high SNR would make the optimal pinhole smaller (to get a sharper image), whereas a low SNR would make the optimal pinhole larger (to get more total light). Indeed, this is exactly what our pupils do! But it's satisfying to see this justified by our information-theoretic model, so let's proceed with that now. Note, also, that for sufficiently small pinholes, the effects of diffraction play an important role [34]. We are ignoring the effects of diffraction in this section. When the SNR is very high, however, a very small pinhole will be optimal, so the results of this section may not apply to very high-SNR regimes.

I will begin by considering not pinholes in the traditional sense, but pinholes in which not only the center of the occluder, but also the edges, transmit light. This implies that the resulting transfer matrix will be circulant, which will make the resulting analysis simpler. For this reason, I will call this kind of occluder frame a "circulant pinhole." In the case of either circulant or traditional pinholes, I will refer to the size of a pinhole by the fraction of the occluder frame that transmits light.

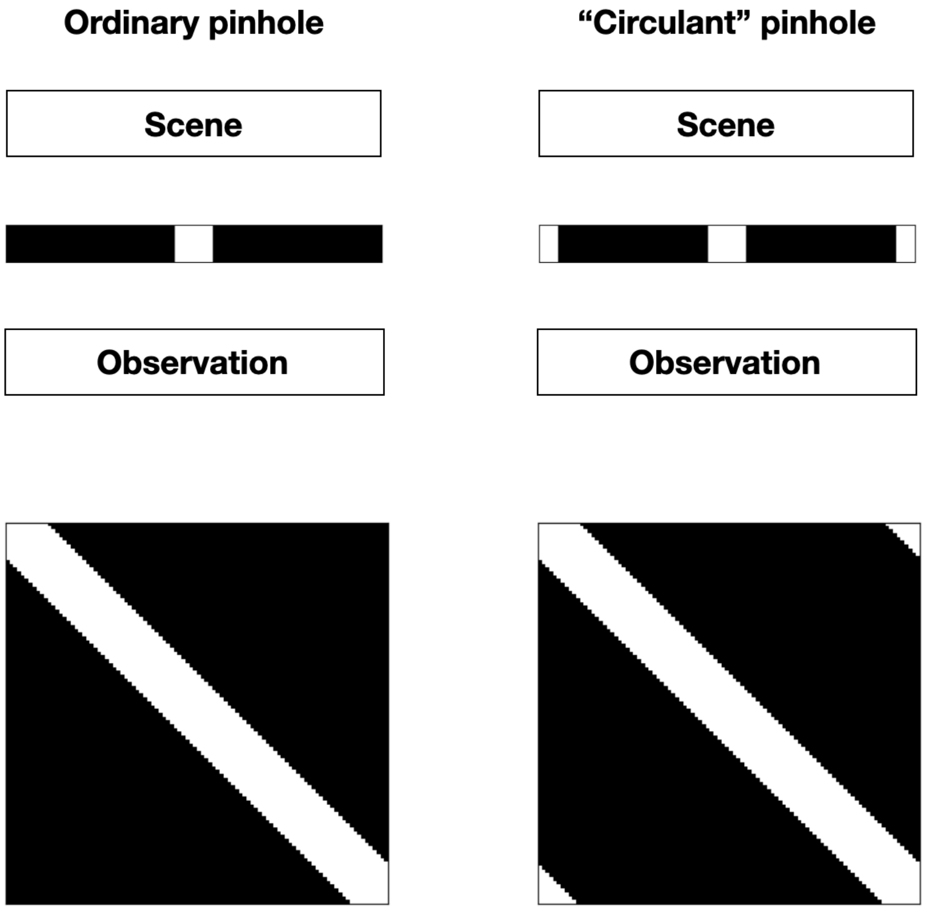


Figure 3-27: Left: a traditional pinhole of size $r = 0.1$, and its associated transfer matrix. Right: a circulant pinhole of size $r = 0.2$, and its associated transfer matrix. Note that the transfer matrix on the left is not circulant, but the one on the right is.

Note that this means that the central hole of a traditional pinhole of size $r = 0.1$ will have the same size as the central hole of a circulant pinhole of size $r = 0.2$, and that a pinhole of size 1, circulant or otherwise, implies an occluder frame with no occlusion at all. See Figure 3-27 for a diagram for what these pinholes may look like.

Consider the formula we considered earlier for the mutual information of any occluder. We have:

$$\mathcal{I} = \log\det\left(\sigma A^T Q(n) A / n^2 + I\right)$$

Here, $Q(n)$ is a the covariance matrix of the scene. It takes the form FDF^*, where D is a diagonal matrix with entries d_i. σ is the signal-to-noise ratio; by default, it's going to be $\theta/(rJ+W)$, where θ is the mean intensity of the scene, J is the "shot noise" or measurement noise, and W is the "ambient noise" or noise from glare. To simplify things, we'll take $J=0$ in this analysis, so that σ is constant in r.

Given that A is circulant, it can be written in the form $F\Lambda F^*$, where Λ is a diagonal matrix containing the eigenvalues of A (equivalently, the frequencies of the first row of A), which I'll refer to as λ_i. It follows that:

$$\mathcal{I} = \log\det\left(\sigma F|\Lambda|^2 DF^*/n^2 + I\right)$$

Using the fact that the eigenvalues of $A+I$ are $\lambda(A)+1$, it follows that the eigenvalues of the matrix $\sigma F|\Lambda|^2DF^*+I$ are $\sigma|\lambda_i|^2d_i/n^2+1$. Because the determinant is the product of the eigenvalues, and the log of a product is a sum, we can write the mutual information as follows:

$$\mathcal{I} = \sum_{j=0}^{n} \log\left(\sigma|\lambda_j|^2 d_j/n^2 + 1\right)$$

Note that so far, we haven't used the fact that the occluder is a pinhole at all, only the fact that it's circulant. Everything so far has been general in the form of the occluder.

But given a specific occluder, we can go further. The eigenvalues λ_j, using the fact that they must be frequencies of the first row of A, a_j, are:

$$\mathcal{I} = \sum_{j=0}^{n} \log\left(\sigma\left|\sum_{k=0}^{n} a_j e^{\frac{-2\pi ijk}{n}}\right|^2 d_j/n^2 + 1\right)$$

In the case of the circulant pinhole, of course, we have $a_j = 1$ for $k \leqslant rn$, and $a_j = 0$ otherwise. (Strictly speaking, this is a rotation of the occluder frame we're considering, but for circulant occluders, this will make no difference.)

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \left| \sum_{k=0}^{rn} e^{\frac{-2\pi ijk}{n}} \right|^2 d_j/n^2 + 1 \right)$$

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \left| \frac{1 - e^{2\pi ijr}}{1 - e^{2\pi ik/n}} \right|^2 d_j/n^2 + 1 \right)$$

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \left| \frac{1 - (\cos(2\pi ijr) + i\sin(2\pi ijr))}{1 - (\cos(2\pi ij/n) + i\sin(2\pi ij/n))} \right|^2 d_j/n^2 + 1 \right)$$

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \frac{(1 - \cos(2\pi ijr))^2 + \sin^2(2\pi ijr)}{(1 - \cos(2\pi ij/n))^2 + \sin^2(2\pi ij/n)} d_j/n^2 + 1 \right)$$

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \frac{2 - 2\cos(2\pi ijr)}{2 - 2\cos(2\pi ij/n)} d_j/n^2 + 1 \right)$$

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \frac{\sin^2(\pi ijr)}{\sin^2(\pi ij/n)} d_j/n^2 + 1 \right)$$

Sadly, this is as far as it's easy to go, without making further assumptions or simplifications. In the limit of finer discretization ($n \to \infty$), the outer sum will become an integral, and we'll need to take an integral resembling $\int \log(\sin^2(x) + 1)$. Unfortunately, this integral doesn't have a closed-form solution, given its complicated dependence on Jonquière's function. Nevertheless, this shows how to perform similar analysis for other kinds of occluders, and we've reduced a complicated expression involving matrices to a single integral.

So what pinhole size *is* optimal, as a function of scene correlation and SNR? Fortunately, with only a single integral, this can be computed quickly. See Figure 3-28 for a plot of the optimal pinhole size as a function of those two parameters. As is apparent, the SNR is of primary importance, with optimal pinhole size having little dependence on scene correlation.

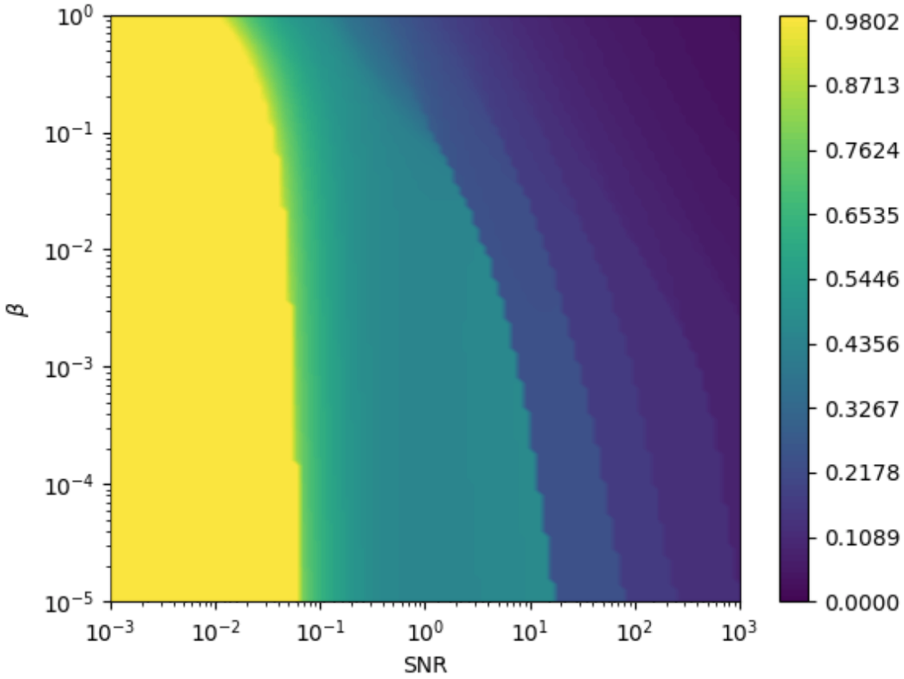


Figure 3-28: Optimal (circulant) pinhole size, as a function of scene correlation and SNR. Note that optimal pinhole size depends primarily on SNR, and only secondarily on scene correlation.

3.14 Szegö's theorem and Toeplitz transfer matrices

We'd like to be able to use techniques like those used in the previous section to analyze Toeplitz transfer matrices, not just circulant ones. In particular, we'd like to be able to express the mutual information directly in terms of the elements of the transfer matrix, as we did earlier:

$$\mathcal{I} = \sum_{j=0}^{n} \log \left(\sigma \left| \sum_{k=0}^{n} a_j e^{\frac{-2\pi ijk}{n}} \right|^2 d_j/n^2 + 1 \right)$$

This was only possible because of the special property of circulant matrices, that the eigenvalues of a circulant matrix are the Fourier transform of any row or column of the matrix. That property does not hold for Toeplitz matrices. Instead, for Toeplitz matrices A, we have Szegö's theorem, which tells us that, for any continuous function $F(\cdot)$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} F(\lambda_j) = \frac{1}{2\pi} \int_0^{2\pi} F(\hat{a}(f))df \tag{3.26}$$

where the λ_j are the eigenvalues of A and

$$\hat{a}(f) = \lim_{n\to\infty} \sum_{k=-n}^{n} t_k e^{ikf}$$

In order for the theorem to hold, we must also have that

$$\sum_{-\infty}^{\infty} |t_k|^2 < \infty \tag{3.27}$$

In Eqs. (3.26) and (3.27), the t_k denote the elements of the first row and first column of the Toeplitz matrix A, with t_{-n} being the bottom element of the first column, t_0 being the top element of the first column (or the first element of the top row), and t_n being the last element of the top row. Given all the t_k from $-n \leqslant k \leqslant n$, we have fully specified the Toeplitz matrix A.

At a very high level, we can think of Szegö's theorem as saying something analogous to what we know about circulant ones: the sum of any function of the eigenvalues of a Toeplitz matrix (the left-hand side of Eq. (3.26)) is equal to the sum of that same function of the frequencies of that Toeplitz matrix's elements (the right hand side of Eq. (3.26)). This is similar to, but not the same as, the eigenvalues of a Toeplitz matrix just being equal to those frequencies, which is what circulant matrices give us.

Note that Szegö's theorem applies only in the $n \to \infty$ limit: the limit of very fine discretization. In fact, it doesn't even apply to specific Toeplitz matrices, only to families of Toeplitz matrices, which must be well-defined as you take $n \to \infty$. Fortunately, that's exactly the case we care about! Our transfer matrices are families of matrices, parametrized by n, the discretization level; and because their every element is 0 or $1/n$, the sum of the first row and column will always be less than or equal to 2. So Szegö's theorem applies in our case.

However, we can't quite use Szegö's theorem as presented. Recall that we are interested in the sum of the logs of the eigenvalues of $\sigma \tilde{A}^T \tilde{A} + I$ (using $\tilde{A} = A/n$, with A a binary matrix, as defined previously). We'd like to relate the sum of the logs of the eigenvalues of $\sigma \tilde{A}^T \tilde{A} + I$ to the elements of $\tilde{A}$, but Szegö's theorem only lets us relate the eigenvalues of $\tilde{A}$ to the elements of $\tilde{A}$. And unfortunately, we can't directly relate the eigenvalues of $\sigma \tilde{A}^T \tilde{A} + I$ to the eigenvalues of $\tilde{A}$. When A is circulant, then $\lambda(\sigma \tilde{A}^T \tilde{A} + I) = |\lambda(\tilde{A})|^2 + 1$, but the same is not true of Toeplitz matrices. (To see why this cannot be true in general at a glance, Toeplitz matrices can be rectangular, and rectangular matrices don't have eigenvalues!) Instead, what we can say is that $\lambda(\sigma \tilde{A}^T \tilde{A} + I) = |s(\tilde{A})|^2 + 1$, where $s(\tilde{A})$ denotes the singular values of $\tilde{A}$.

So now we can relate the eigenvalues of $\sigma \tilde{A}^T \tilde{A} + I$, which is what we care about, to the singular values of $\tilde{A}$. Can we relate the singular values of $\tilde{A}$ to the elements of $\tilde{A}$? Fortunately, we can! Thanks to an extension of Szegö's theorem due to S. Parter [79], we have:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} F(s_j) = \frac{1}{2\pi} \int_0^{2\pi} F(|\hat{a}(f)|) df \tag{3.28}$$

where s_j is the j^{th} singular value of $\tilde{A}$. Note that this equation is nearly identical to Equation (3.26), with the only difference being that we take the norm of each frequency before feeding it to the function F, on the right-hand side of the equation.

Now that we have this, how do we use it to compute the mutual information we'd get from a given Toeplitz transfer matrix? We know that the mutual information of a configuration with transfer matrix $\tilde{A}$, an SNR of σ, and an IID scene distribution is given by:

$$\sum_{j=0}^{n-1} \log(\sigma |s_j|^2 + 1)$$

(Note that because we're using $\tilde{A}$ and not A, there's no factor of $1/n^2$ in the formula.)

This means that the function F we're interested in is defined as:

$$F(x) = \log(\sigma |x|^2 + 1).$$

Thus, thanks to Parter's extension of Szegö's theorem, we can cleanly express the mutual information directly in terms of the elements of the transfer matrix, at least under certain conditions. We have:

$$\lim_{n\to\infty} \mathcal{I} = \frac{1}{2\pi} \int_0^{2\pi} \log(\sigma |\hat{a}(f)|^2 + 1) df$$

where, as before,

$$\hat{a}(f) = \lim_{n\to\infty} \sum_{k=-n}^{n} t_k e^{ikf}$$

and the t_k are the elements of the first row and first column of the Toeplitz matrix.

Szegö's theorem is a nice tool to have for studying Toeplitz transfer matrices. The best way to use it may be to find covariance matrices $\mathbf{Q}$ that will happen to be diagonalized by the same matrices that diagonalize $\tilde{A}$, in which case we will be able to find coherent families of Toeplitz matrices to analyze.

3.15 The tomography model and the light-field model, compared

Planar scenes are very convenient to represent and analyze, which is why so much of this thesis considers them in particular. Moreover, suppose we are already taking the paraxial approximation, which, as described in Section 1.3, is fairly realistic for scenes that are twice as far from the observation as the observation is wide. In those cases, assuming a non-planar scene to be planar produces no *additional* distortion over that introduced by the paraxial approximation already, besides, of course, the fact that you're only getting a 2D view of a 3D object. This isn't as bad as it sounds, since after all, a photograph is only a 2D view of a 3D object! Despite that, we're usually able to get a good sense of what a scene looks like from a photograph.

So in fact, in general, assuming scenes are planar won't deny you that much information about the scene, if you can get a high-quality 2D reconstruction. That said, there are certainly applications where depth is the main thing you care about, foremost among them coded-aperture cameras, one of the applications of which is that they give you depth information that an ordinary lens-based camera wouldn't; see for example [66]. In those cases, assuming the scene is planar will defeat the whole purpose of what you're doing, and because coded apertures are an important application of the occluder-based imaging model that I'm presenting here, that's a big problem. So it's important that we be able to remove the assumption that scenes are planar and still be able to apply the transfer-matrix-based style of analysis we're used to.

Thus the question: how do we represent the scene, which is now a 3D object? The easiest way to do it while preserving linearity, which is crucial in order for the transfer matrix to make sense as a representation of the action of the occluder (see Subsec. 1.4.1 for a refresher), is to represent the scene as a light-field. A "light-field" here is defined as a 2D view of the scene, as seen from each point on a 2D sensor array. See Figure 3-29 for an example of what a perfect light field reconstruction would look like in a real-world setting.

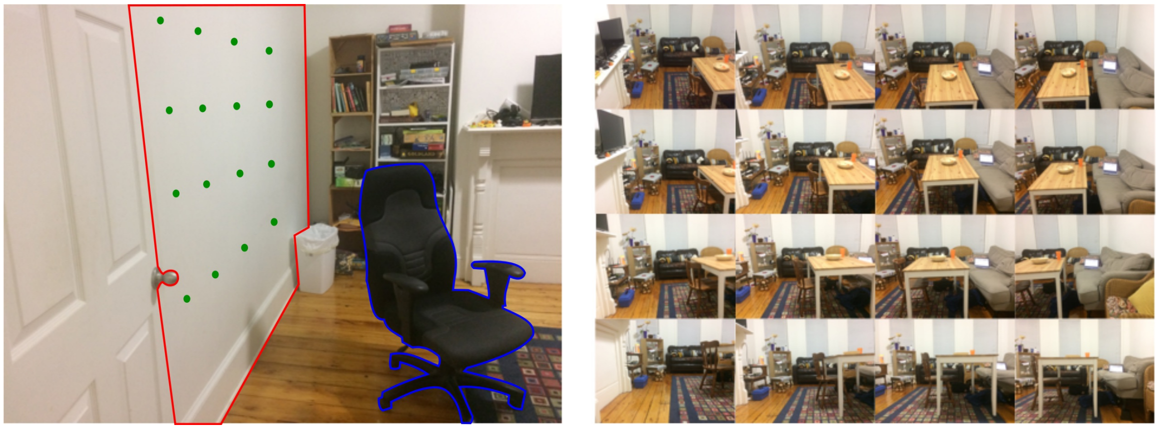


Figure 3-29: If you could perfectly reconstruct a light field using an occluder (the chair, outlined in blue) the reconstruction would be what can be seen on the right of this figure: the scene, as seen from every point (in green) on the observation plane (outlined in red).

A light field is more than enough to represent an entire 3D scene; after all, if you can see the scene from every point on the observation plane, you know everything there is to know about the scene (or at least, everything that could possibly be learned by studying the observation plane). And the notion of the transfer matrix is easily extended to a light-field; just put a 1 in every entry of the transfer matrix corresponding to an scene-point/observation-point pair with an unoccluded view of each other, and a 0 elsewhere. (See Fig. 4-16)

So what's wrong with a light-field as a means of representing a 3D scene? The problem is that it's wasteful. Assuming the scene is $n \times n \times n$, it should only require n^3 vector entries to represent it; yet an $(n \times n) \times (n \times n)$ light field will require n^4 entries to represent it.[4] The scene could have been more parsimoniously represented, and as a result the reconstruction algorithm will take longer to run than necessary (because the transfer matrix will be bigger than it needed to be). Fortunately, the dependence on the transfer matrix size will only be linear; the algorithmic bottleneck will be in computing AQA^T when A is an $n^2 \times n^4$ matrix. This takes $O(n^6)$ time, when if we'd represented the scene as a length-n^3 vector, it would have taken only $O(n^5)$ time.

[4] Note that this tradeoff doesn't exist in flatland; both a voxelized scene with depth and a light-

The other obvious way to represent a 3D scene is to voxelize it and treat it as a 3D array of light sources. This works perfectly if the application is tomography, where there's no or little self-occlusion within the 3D scene space (hence the term "tomography model"); however, if there's substantial self-occlusion within the scene, then this approach offers us no way to model it. Occlusion between different scene elements, after all, behaves nonlinearly with respect to the amount of light hitting the observation plane. To explain what I mean by this, I'll give a concrete example. Suppose that we are interested in one point in the observation plane, whose intensity we'll call $c(a, b)$. Suppose now that there are two possible light sources in the scene, a and b, which can be 0 or 1, and that b occludes a. Linear behavior would be if (for example) $c(a, b) = a + b$. But if b occludes a, then instead what we have is $c(a, b) = b$ if $b = 1$, and $c(a, b) = a$ otherwise. This is nonlinear!

The least-squares 3D non-self-occluding representation of a 4D light-field can be obtained using the $n^4 \times n^3$ matrix that converts 3D non-self-occluding spaces to 4D light-fields. Such a matrix exists for non-self-occluding spaces, but not for self-occluding ones, since only the former spaces behave linearly. If we call this matrix M, we can generate the least-squares estimate for which 3D non-self-occluding space would generate a given light-field l by computing $(M^T M + \lambda I)^{-1} M^T l$, where λ is a regularization term. Figure 3-31 was generated with this method using $\lambda = 10^{-3}$.

How much distortion is introduced by simply ignoring the nonlinearity? Quite a lot, depending on what representation you use at the end. In Figure 3-30 we have a simulated three-dimensional scene with two frames, the red one partially occluding the blue one. If we take the light-field that results and try to estimate a non-self-occluding 3D space that generates it (which is impossible, of course, since the true scene *does* self-occlude), the least-squares solution 3D space looks like complete nonsense (see Figure 3-31). However, the light-field representation of the nonsense 3D space does indeed come very close to the true light-field! This is one example where the wrong

field will require n^2 vector entries to represent, making the light-field representation strictly better. In general, representing a d-dimensional space in a d-dimensional world will require n^d entreis to represent the traditional way, and $n^{2(d-1)}$ entries to represent using a light-field, making light-fields less and less efficient as d grows.

model is, unfortunately, not useful.

3.16 Non-planar occluders and scenes

This section explores what happens when we relax the requirement that occluders and scenes be planar. Under such circumstances, which occluders are optimal?

The easiest case to consider is the case where the scene is still presumed to be planar, but the occluder need not be. In this case, an optimal planar occluder still exists, since a spectrally flat occluder halfway in between the scene and observation plane achieves Hadamard's bound, and all transfer matrices in this case will still be square binary matrices.

If we relax the planarity requirement for both the scene and the occluder, analysis once again becomes very challenging, for many of the same reasons as described in Section 3.11. If we give the scene depth, the transfer matrix will be the horizontal concatenation of several different transfer submatrices, with each one corresponding to a planar scene at that depth (see for instance the bottom of Figure 3-32). If we call each of these transfer submatrices A_i, the overall transfer matrix will be given as $A = \{A_1, \ldots, A_k\}$, where k is the number of discrete depths in the scene we are considering. Ignoring scene correlations, we can then express the mutual information in terms of a sum over the submatrices:

$$\mathcal{I} = \log\det\left(\sigma\frac{AA^T}{n^2} + I\right) = \log\det\left(\sigma\frac{1}{n^2}\sum_{i=1}^{k} A_i A^T + I\right)$$

Unfortunately, there's not very much further you can go from here, that I know of; there's not much to be said that's known about the determinant of a sum of matrices, even if, as in this case, those matrices are known to be positive semidefinite.

That said, it would be nice to gather empirical evidence for or against the conjecture that planar occluders are optimal even for non-planar scenes. To do this, I performed an exhaustive and a greedy search, shown in Figs 3-32 and 3-33, respectively, over binary occluders at multiple depths. If the optimizations returned

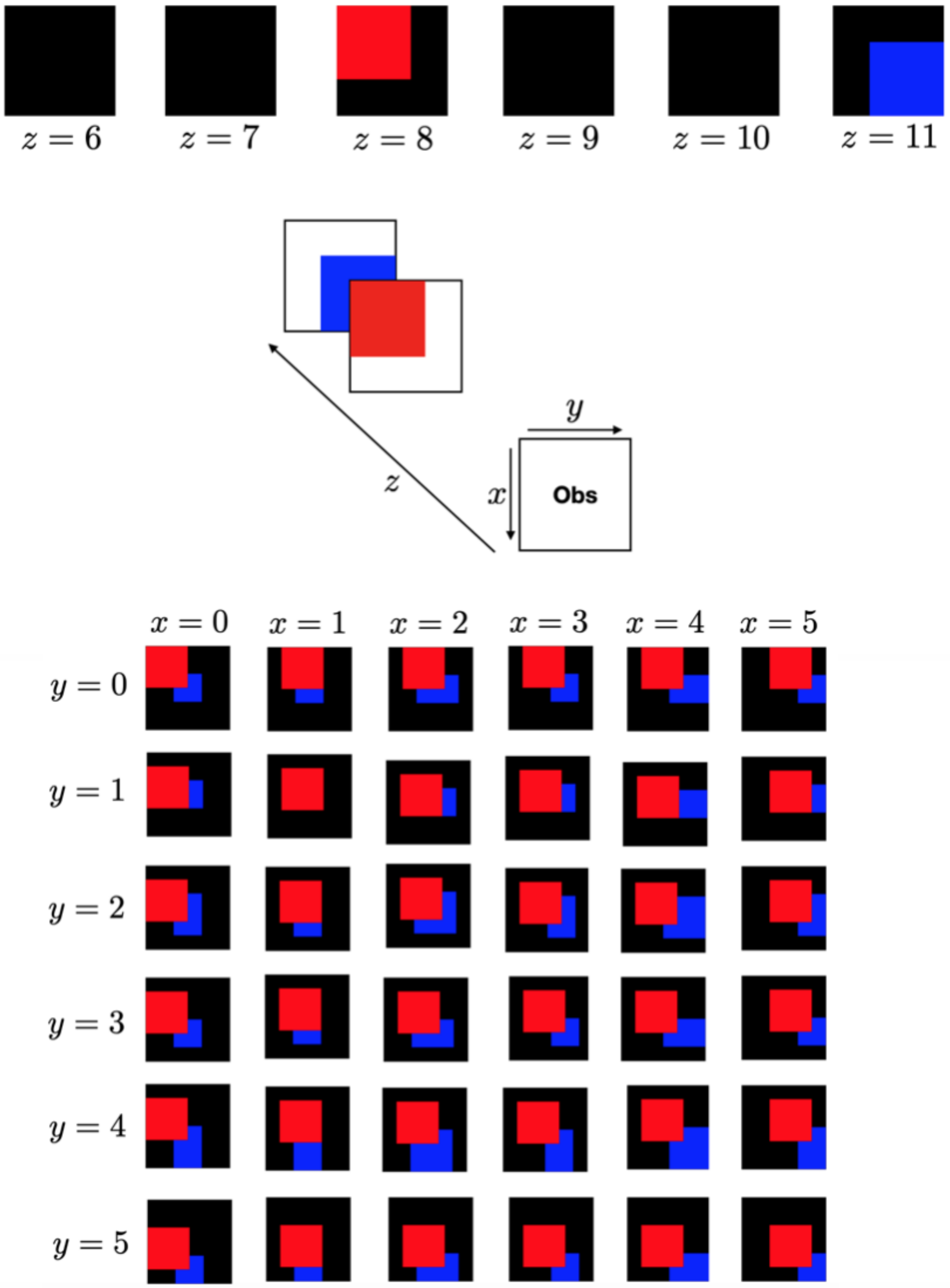


Figure 3-30: Top: the configuration of the imaging system. The scene is made up of two opaque frames, one red and one blue, at $z = 8$ and $z = 11$, respectively. Bottom: the light-field corresponding to that configuration (i.e. the appearance of the scene, viewed from every point on the observation plane).

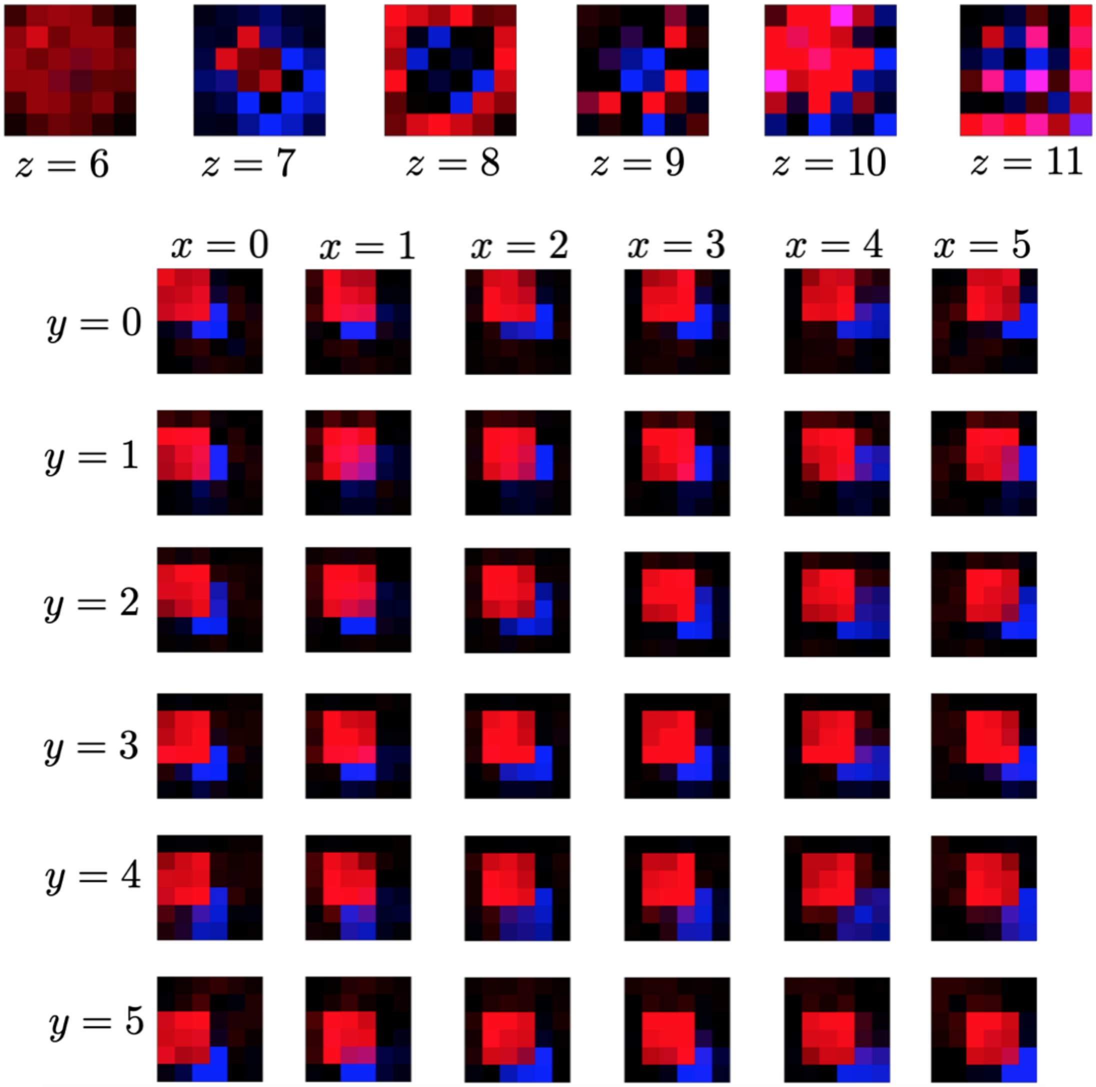


Figure 3-31: Top: the estimated non-self-occluding 3D space, using a least-squares method. Bottom: the light-field derived from that space. As can be seen, despite the fact that the 3D space estimate is a terrible estiamte of the true 3D space, the light-field that results is very close to the true light-field.

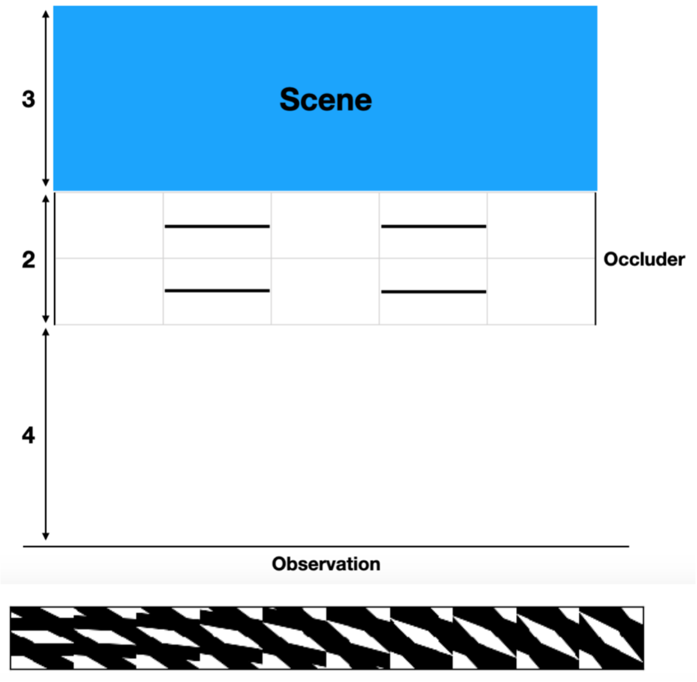


Figure 3-32: Top: the scene configuration. Note that we are using the paraxial approximation, so the width of the scene and observation plane are taken to be much smaller than the scale of this figure would suggest. The optimal occluder (found by exhaustive search over all 2^{10} occlusion possibilities in the 2×5 grid shown in grey) is not planar. The scene is IID, $\sigma = 10^3$, $k = 10$, $n = 100$. Bottom: the $n \times kn$ transfer matrix corresponding to the occluder shown above.

occluders at a single depth despite having the option of putting occlusion at multiple different depths, that would be some evidence in favor of the conjecture; if not, that would be strong evidence against it. Both optimizations returned occluders with occlusion at multiple different depths, strongly suggesting that optimal occluders for non-planar scenes are not planar either.

3.17 Extensions into the 3D world

Throughout this chapter, the default model analyzed has been a “flatland”—i.e. a 2D world—model. I’ve alluded on several occasions to generalizations of the model to a 3D world, and I keep the bulk of my analysis in 2D because it’s both the case that those generalizations are generally intuitive and straightforward, and that keeping things in 2D helps avoid unnecessary complexity, and makes situations much easier

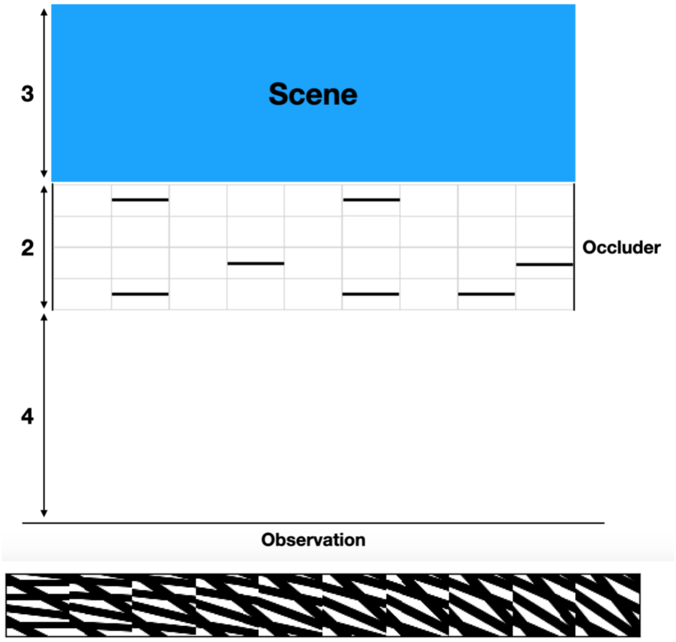


Figure 3-33: Top: the scene configuration. Note that we are using the paraxial approximation, so the width of the scene and observation plane are taken to be much smaller than the scale of this figure would suggest. The locally optimal occluder (found by greedy search over the 2^{36} occlusion possibilities in the 4×9 grid shown in grey) is not planar. The scene is IID, $\sigma = 10^3$, $k = 10$, $n = 99$. Bottom: the $n \times kn$ transfer matrix corresponding to the occluder shown above.

to visualize on a 2D page. That said, my analysis would not be complete if I did not make explicit how to generalize from 2D to 3D.

First of all, as previously mentioned, 2D analysis can be useful when integrating over one dimension of space. Why would one want to integrate over one dimension? There are a variety of reasons. One is to increase the SNR of the problem, since you are increasing your signal strength by a factor of n, if you are integrating along one dimension of an $n \times n$ observation plane. Another is because the dimension you're integrating along doesn't have much information contained in it anyway; for an example of a problem with that property, see Section 4.1. When looking at the shadow cast by a building's corner, it's not that the variation along radial lines outward from the corner contains *no* information about the hidden scene; it's that the amount of information along the r direction is far less than that along the θ direction, so you're better served by integrating along the less informative direction.

What about cases where both dimensions are equally informative, though? In such cases, it's useful to be able to generalize cleanly from 2D to 3D. Fortunately, in most cases the generalization is very simple. I'll go through each case individually.

3.17.1 The illumination function in 3D

As explained in Section 1.3, in a 2D world, the illumination function of a flat surface at $y = 0$ by a point light source at $(0, y_p)$ is given by:

$$I(x) = \frac{y_p}{2\pi(x^2 + y_p^2)}$$

whereas in a 3D world, the illumination function of a flat surface at $z = 0$ by a point light source at $(0, 0, z_p)$ is given by:

$$I(x, y) = \frac{z_p}{4\pi(x^2 + y^2 + z_p^2)^{3/2}}$$

3.17.2 Spectrally-flat occluders in 3D

Note that this generalization is still assuming a flat occluder, if not a flat world; that means that if the world is going from 2D to 3D, the occluder is going from 1D to 2D.

One simple and intuitive generalization is: given two spectrally-flat 1D occluders a and b, you can make a spectrally-flat 2D occluder C by taking the XOR outer product:

$$C_{ij} = a_i \oplus b_j,$$

where $\oplus$ denotes XOR. Here I mean "spectrally-flat" in the non-trivial sense described in Section 3.2.2. Given that, let's suppose that a has n elements and b has m elements. All the frequencies of a have a magnitude of $\sqrt{n+1}/2$, except for the DC term, which has a magnitude of $(n+1)/2$. Similarly, all the frequencies of b have a magnitude of $\sqrt{m+1}/2$, except for the DC term, which has a magnitude of $(m+1)/2$.

It's easy to check that C, as defined above, will also be spectrally flat in a similar sense. Taking the Fourier transform of C to be $\bar{C}$, the magnitude of the DC-DC term, $|\bar{C}_{00}|$, is given by:

$$|\bar{C}_{00}| = (mn + 1)$$

The magnitudes of the DC-AC and AC-DC terms, $|\bar{C}_{0j}|$ and $|\bar{C}_{i0}|$ (assuming $i > 0$ and $j > 0$), are given by:

$$|\bar{C}_{0j}| = \sqrt{m+1}/2$$

$$|\bar{C}_{i0}| = \sqrt{n+1}/2$$

Finally, the magnitudes of the bulk of the frequencies, the AC-AC terms, are indeed flat (this is the defining feature of a 2D spectrally-flat occluder):

$$|\bar{C}_{ij}| = \sqrt{(m+1)(n+1)}/2$$

So we maintain this desirable property of spectral flatness from 1D to 2 via a simple outer product.

It is important to note that while this is useful construction to create spectrally-flat 2D occluders, it is by no means, the *only* construction; others exist as well that do not have the property that they are the XOR outer product of two 1D spectrally flat patterns. For an example of such, see Figure 2-2.

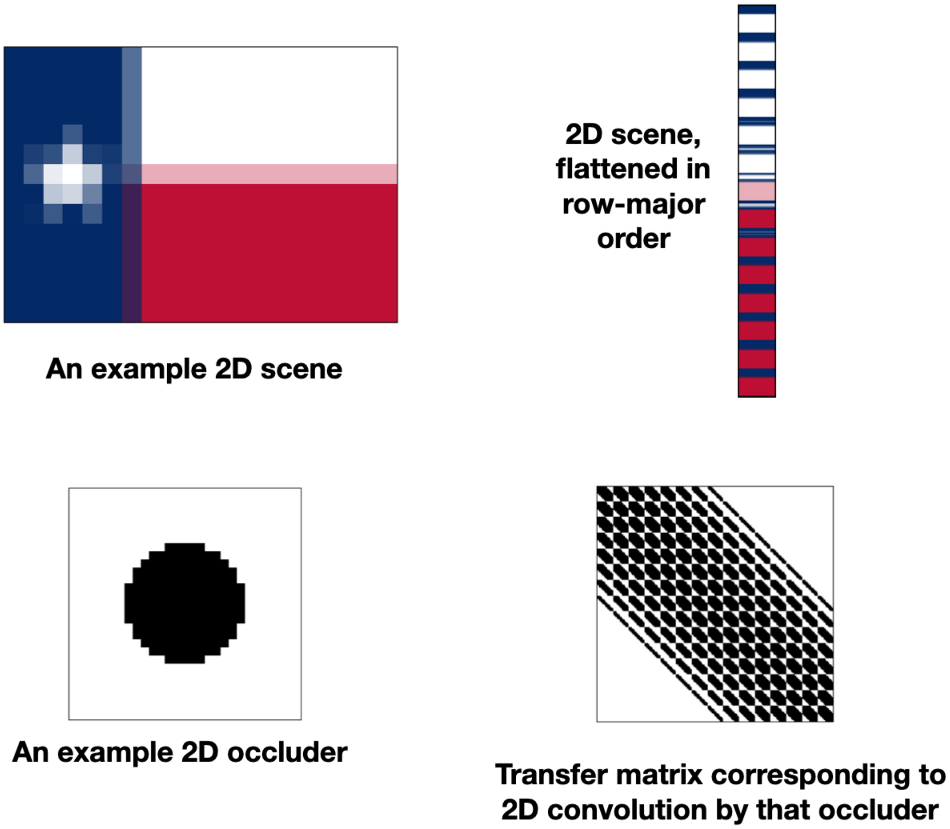
3.17.3 Convolutional transfer matrices in 3D

I've described in Section 1.4.1 why it is that transfer matrices corresponding to occluders should be Toeplitz, or, when they repeat themselves once, circulant (see Figure 3-1 if you've forgotten why that follows). It's natural to ask the question: what about transfer matrices corresponding to 2D occluders? What do they look like?

Exactly what they look like is a question of notation. Throughout this thesis, I generally assume that 2D scenes and observations are being represented by vectors. Of course, representing a 2D object with a vector, rather than a matrix, requires you to flatten the object somehow; you have to represent the image using a vector using (for example) row-major order.

This is fine, but if this is the notation style you use, a convolutional transfer matrix—meaning one that represents the action of a 2D occluder—won't be circulant or Toeplitz in the traditional sense. Rather, just as a circulant matrix is diagonalized by the Fourier basis, a "2D-circulant" matrix will be diagonalized by the flattened Fourier basis, and multiplication by a 2D-convolutional matrix will be equivalent to a 2D convolution. Rather than multiplying by such a matrix, it will be much more computationally efficient to just perform the convolution, for exactly the same reasons as in 1D.

In the future, when I reference convolutional transfer matrices corresponding to 2D occluders, I'll be referring to matrices that have the property of being equivalent to a convolution in 2D, and being diagonalized or near-diagonalized by the 2D Fourier basis. See Figure 3-34 for an diagram of matrices with this property.



An example 2D scene
2D scene, flattened in row-major order
An example 2D occluder
Transfer matrix corresponding to 2D convolution by that occluder

Figure 3-34: An example scene flattened using row-major order, and an example 2D occluder's corresponding convolutional transfer matrix. Note that the transfer matrix is "2D-Toeplitz" without being Toeplitz in the ordinary sense. This means that it will be near-diagonalized by the flattened 2D Fourier basis (rather than the 1D Fourier basis).

3.17.4 Tensor notation for convolutional transfer matrices

An alternate and equivalent way to represent convolutional transfer matrices in 3D is to use matrices to represent the scene and observation planes, and to have a transfer tensor represent the action of the occluder. For example, suppose we had a $x_s \times y_s$ scene X, and a $x_o \times y_o$ observation plane Y. We could, as described in the previous section, flatten the scene and observation and represent them as length-$x_s y_s$ and length-$x_o y_o$ vectors, respectively, and the transfer matrix as a $x_o y_o \times x_s y_s$ matrix. Alternately, we could formulate the action of the occluder as a $(x_o \times y_o) \times (x_s \times y_s)$ *tensor* $\mathcal{A}$.

The easiest way to think about $\mathcal{A}$ is as a $x_o \times y_o$ matrix, where every element of that matrix is *itself* a $x_s \times y_s$ matrix. Then, the product of $\mathcal{A}$ and a $x_s \times y_s$ matrix (such as Y, the matrix that represents the scene) is by returning a $x_o \times y_o$ matrix, whose every element is the Frobenius inner product of each element of $\mathcal{A}$ with X.

The Frobenius inner product is the product that takes in two matrices with identical shapes, and returns a scalar. It's the sum of the elementwise product of the two matrices; informally speaking, it's the "dot product" of the two matrices. From now on, I'll use the $\oplus$ operation to denote it. $U \oplus V = tr(U^T V) = \sum_{ij} U_{ij} V_{ij}$.

So more formally, the product of the $(x_o \times y_o) \times (x_s \times y_s)$ transfer tensor $\mathcal{A}$ and a $x_s \times y_s$ matrix X is defined as follows:

$$(\mathcal{A}X)_{ij} = \mathcal{A}_{ij} \oplus Y, 1 \leqslant i \leqslant x_o, 1 \leqslant j \leqslant y_o$$

See Figure 3-35 for a worked example of tensor multiplication as defined here.

Within this framework, a circulant transfer tensor is a straightforward extension of the notion of a circulant transfer matrix. Submatrices in the transfer tensor are analogous to rows in the transfer matrix, so horizontally adjacent submatrices of a circulant transfer tensor are horizontal shifts of one another, and ditto with vertically adjacent transfer matrices. See Figure 3-36 for an example of a circulant transfer tensor.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} E = \begin{bmatrix} A\cdot E & B\cdot E \\ C\cdot E & D\cdot E \end{bmatrix} \qquad A\cdot E = \mathrm{tr}(A^T E)$$

(aka elementwise dot product of entries)

$$\left(\begin{matrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix} \right) \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 7 & 4 \end{bmatrix}$$

Figure 3-35: This section's definition of tensor-matrix multiplication, with a worked example.

$$\left(\begin{matrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} & \begin{bmatrix} b & c & a \\ e & f & d \\ h & i & g \end{bmatrix} & \begin{bmatrix} c & a & b \\ f & d & e \\ i & g & h \end{bmatrix} \\ \begin{bmatrix} d & e & f \\ g & h & i \\ a & b & c \end{bmatrix} & \begin{bmatrix} e & f & d \\ h & i & g \\ b & c & a \end{bmatrix} & \begin{bmatrix} f & d & e \\ i & g & h \\ c & a & b \end{bmatrix} \\ \begin{bmatrix} g & h & i \\ a & b & c \\ d & e & f \end{bmatrix} & \begin{bmatrix} h & i & g \\ b & c & a \\ e & f & d \end{bmatrix} & \begin{bmatrix} i & g & h \\ c & a & b \\ f & d & e \end{bmatrix} \end{matrix} \right)$$

Figure 3-36: A circulant $(3 \times 3) \times (3 \times 3)$ tensor.

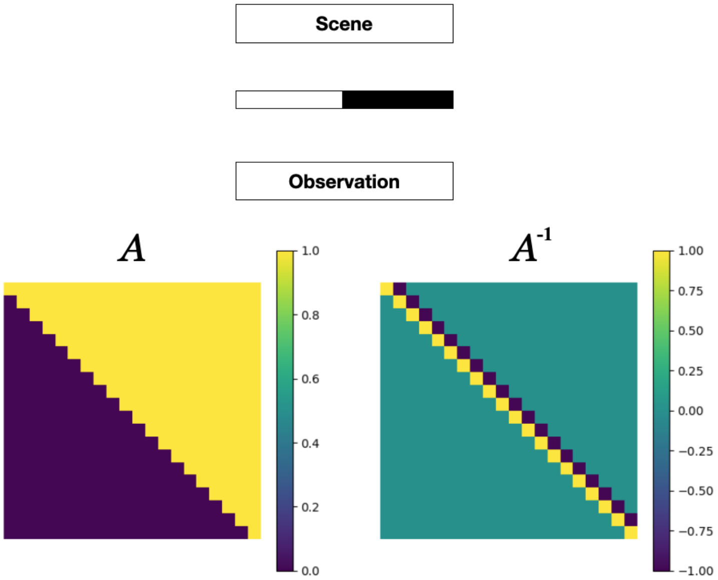


Figure 3-37: Top: a flatland configuration with an edge occluder. Bottom left: the transfer matrix corresponding to this occluder. Bottom right: the inverse of that transfer matrix. The transfer matrix A corresponds to spatial integration, and its inverse A^{-1} to spatial differentiation.

3.18 Separable Occluders

Certain 2D occluders have a special property, which I refer to as *separability*. The action of such occluders can be described in terms of two ordinary "flatland" operations.

Before defining what I mean formally, I'll give a simple example, because I think it helps a lot to clarify this concept. Consider an "edge" occluder in flatland, meaning an occluder whose left half transmits light, and whose right half occludes it. It follows that this occluder's corresponding transfer matrix is the upper-triangular matrix of 1's.

The upper-triangular matrix of 1's is the discrete integral operator; it integrates the scene to produce the observation. It follows, of course, that the inverse matrix is the discrete spatial derivative operator (See Figure 3-37). (This insight is crucial to imaging system described in Section 4.1.)

What's the equivalent outside of flatland? The equivalent 2D occluder is an occluder divided into four quadrants, where one of the four quadrants transmits light, and the other three occlude it. Happily, this is a pattern that shows up a lot in real life, such as in doorways or near windows. (See Figure 3-38) The action of such an occluder is to integrate the scene along both spatial dimensions. To invert it, it suffices to differentiate along both spatial dimensions.

This particular occluder has an additional wonderful property. Because integration along both spatial dimensions is commutative in which spatial dimension you integrate along first (and ditto for differentiation), this occluder is *separable.* This means that in order to recover the scene from the observation, rather than multiplying the observation matrix by the inverse of the transfer tensor, it suffices to multiply the observation matrix by two matrices; one to invert the integration in the x-coordinate, and the other to invert the integration in the y-coordinate. In other words, rather than needing to do $X = \mathcal{A}^{-1}Y$ (which requires an expensive operation on the full $(x_o \times y_o) \times (x_s \times y_s)$ transfer tensor), it suffices instead to do $X = U^{-1}Y(V^{-1})^T$, for particular matrices U^{-1} and V^{-1}. In this particular instances, those matrices U^{-1} and V^{-1} are both equal to the discrete derivative matrix (see Figure 3-37 if you've forgotten what that looks like); multiplying the observation by the same discrete derivative matrix on right and left means taking the derivative along the x- and y-coordinates.

So what exactly does it mean for a $(x_o \times y_o) \times (x_s \times y_s)$ transfer tensor $\mathcal{A}$ to be separable? It means that:

$$\mathcal{A}X = UXV^T$$

for all $x_s \times y_s$ scenes X, and for some particular co-transfer matrices U and V, of shapes $x_o \times x_s$ and $y_o \times y_s$, respectively. And given that $\mathcal{A}$ is separable, the same must, of course, work for its inverse:

$$\mathcal{A}^{-1}Y = U^{-1}Y(V^{-1})^T.$$

And happily, if the transfer tensor $\mathcal{A}$ is separable, and circulant or Toeplitz, the

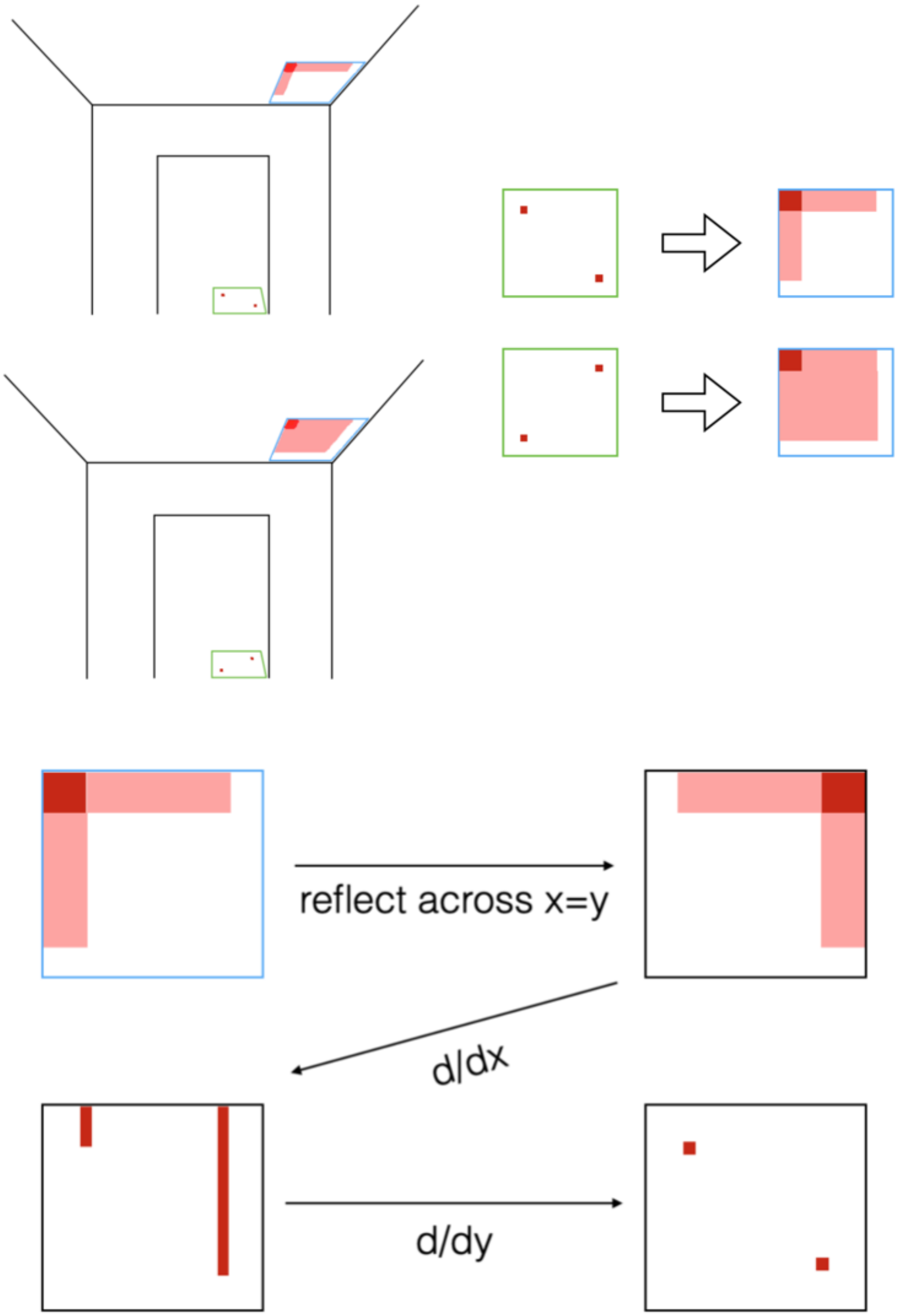


Figure 3-38: Top: two imaging configurations near a doorway; the two different scenes lead to two different light reflections on the ceiling. Note that the observation on the ceiling is equal to the scene, integrated along both spatial dimensions. Bottom: the inversion process for this occluder. Because this occluder is separable, it suffices to do two operations, one along one of the dimensions, and one along the other.

matrices U and V will also be circulant or Toeplitz (as well their inverses).

A transfer tensor being separable obviously makes reconstruction much faster, since matrix multiplication is much faster than tensor-matrix multiplication. Just to give a basic sense of how much faster, if we take $x_o = y_o = x_s = y_s = n$, tensor-matrix multiplication will take $\Theta(n^4)$ time, whereas the reconstruction using separable occluders will take $\Theta(n^\Omega)$ time, where Ω is the matrix multiplication runtime exponent (somewhere between 2 and 3).

We can call occluders whose transfer matrices will be separable "separable occluders." Which occluders are separable? It turns out that it is very straightforward to identify them. An occluder is separable if, when represented with a matrix M of 0's and 1's (with 1's representing the transmissive elements of the occluder, and 0's representing the occluding elements), M is rank-1. (Note that M is *not* a transfer matrix!)

How can we find the matrices U and V that do the same thing as multiplying by the transfer tensor $\mathcal{A}$? Thankfully, that, too, is easy. If $\mathcal{A}$ has the property that each of its submatrices is rank-1 (which would follow from being a transfer tensor representing a rank-1 occluder), then we can write each sub-matrix of $\mathcal{A}$ as the outer product of two vectors, of length x_s and y_s, respectively. Each of those vectors is drawn from two sets of vectors, of sizes x_o and y_o, respectively.

$$\mathcal{A}_{ij} = u_i v_j^T, 1 \leqslant i \leqslant x_o, 1 \leqslant j \leqslant y_o.$$

It is easy to guess what U and V must be: they are the sets of vectors $\{u_1, \ldots, u_o\}$ and $\{v_1, \ldots, v_o\}$, respectively, stacked on top of each other! See Figure 3-39 for a worked example.

3.19 Equivalent Occluders

This section introduces the concept of equivalent occluders. Equivalent occluders are a phenomenon whereby certain transfer matrices or tensors can be substituted for others while still accurately reconstructing the scene. The main use I have found

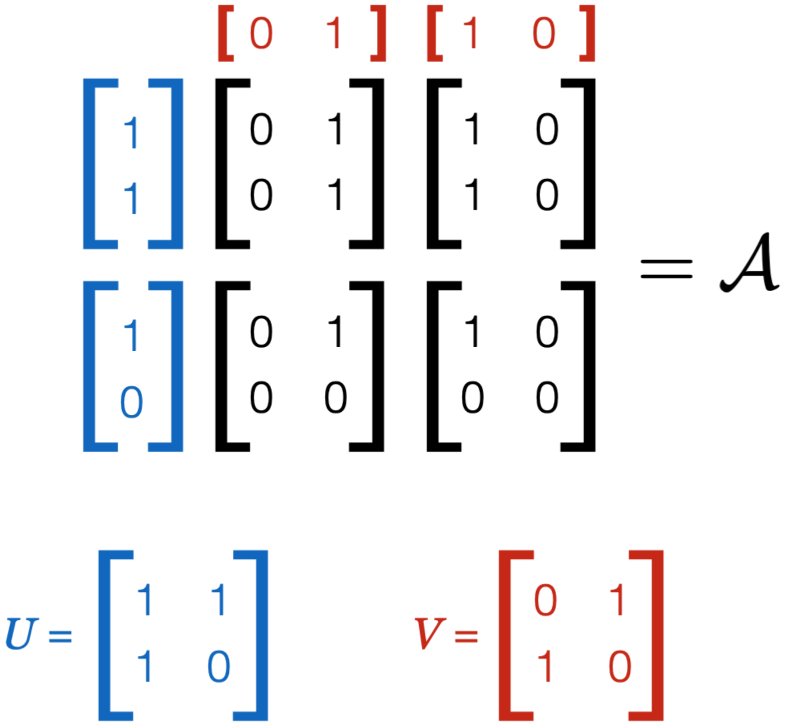


Figure 3-39: A $(2 \times 2) \times (2 \times 2)$ separable transfer tensor $\mathcal{A}$ (in black). Each of its submatrices can be expressed as an outer product of vectors u_i (blue) and v_i (red), each of which make up the matrices U and V respectively. For all 2×2 matrices X, we will have $\mathcal{A}X = UXV^T$.

for this phenomenon is substituting a separable occluder for a non-separable one (that is, inverting a separable transfer tensor that is equivalent to the "real" transfer tensor). In particular, an occluder frame that is an "anti-doorway"—one divided into quadrants where one of the quadrants occludes light while the other three transmit—is not separable, but is equivalent to the doorway occluder from Figure 3-38, which is. This lets you use the techniques from the previous section to invert, which can be tremendously convenient, since this is a common kind of occluder in the real world (such as in the case of using furniture as an occluder).

Beyond that, I'm not sure yet what the applications of families of equivalent occluders might be, but it seems that it could be potentially useful.

To give an example, I'll start with the pair of equivalent occluders that I described earlier: the right-angle occluder (one quadrant off, three quadrants on), and the right-angle non-occluder (one quadrant on, three quadrants off). Let's call the former A_1, and the latter A_2. Suppose that in reality, the occluder frame is represented by the transfer matrix (or tensor, if you prefer that notation) A_1, so that $Y = A_1 X$. We know from the previous section, though, that because A_2 is separable and A_1 isn't, that $A_2^{-1} Y$ will be easier to compute than $A_1^{-1} X$. What happens if we try to reconstruct X by calculating $A_2^{-1} Y$, even though $Y = A_1 X$? In other words, if we let $\hat{X} = A_2^{-1} A_1 X$, how close will X be to $\hat{X}$?

As we can see from Figure 3-40, surprisingly close, for this particular choice of A_1 and A_2. All the pixels not on the border of the reconstructed image agree exactly. This is very strange! The two potential observations that we'd get from multiplying X from each of A_1 and A_2 look completely different from each other. But it turns out that pairs of occluder frames M_1 and M_2 such that $\frac{d}{dx}\frac{d}{dy} M_1 = \frac{d}{dx}\frac{d}{dy} M_2$ have this interesting and occasionally useful property, of being able to freely substitute one for the other while doing almost no damage to the reconstruction you get.

Not *all* the reconstructed pixels are exactly correct, though. The pixels at the edge of the reconstruction $\hat{X} = A_2^{-1} A_1 X$ are wrong; and not just that—they're *very* wrong! In fact, the magnitude of the pixels at the edge of the image is greater than the pixels not at the edge of the image by a factor of the total number of pixels in

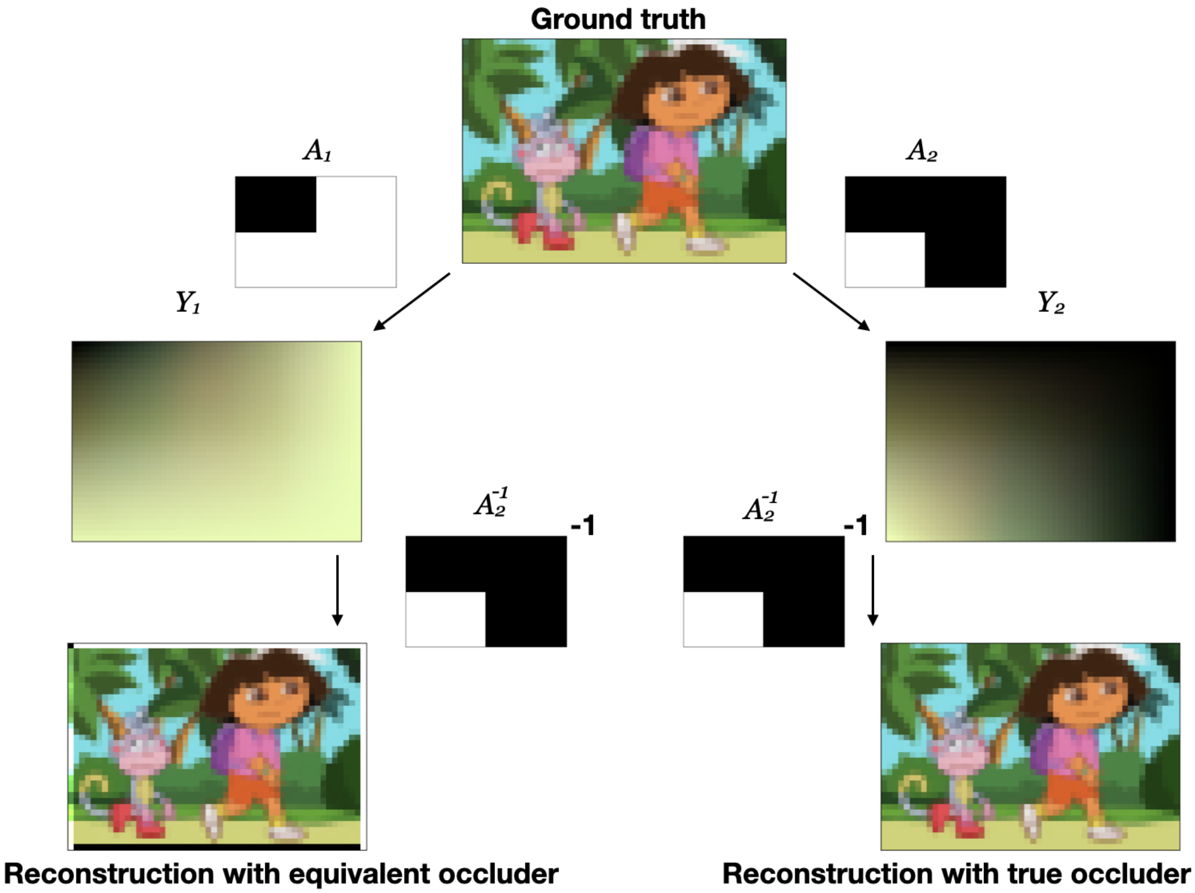

Ground truth
A₁
A₂
Y₁
Y₂
A₂⁻¹
-1
A₂⁻¹
-1
Reconstruction with equivalent occluder
Reconstruction with true occluder

Figure 3-40: A simulated scene, transformed by two equivalent occluders, A_1 and A_2. Despite the fact that the two potential observations Y_1 and Y_2 you would get by transforming the scene X by each of A_1 and A_2 look nothing alike, the reconstruction you get by computing $A_2^{-1}Y_1$ is almost identical to the scene. This is useful, because A_2 is a separable occluder and A_1 is not.

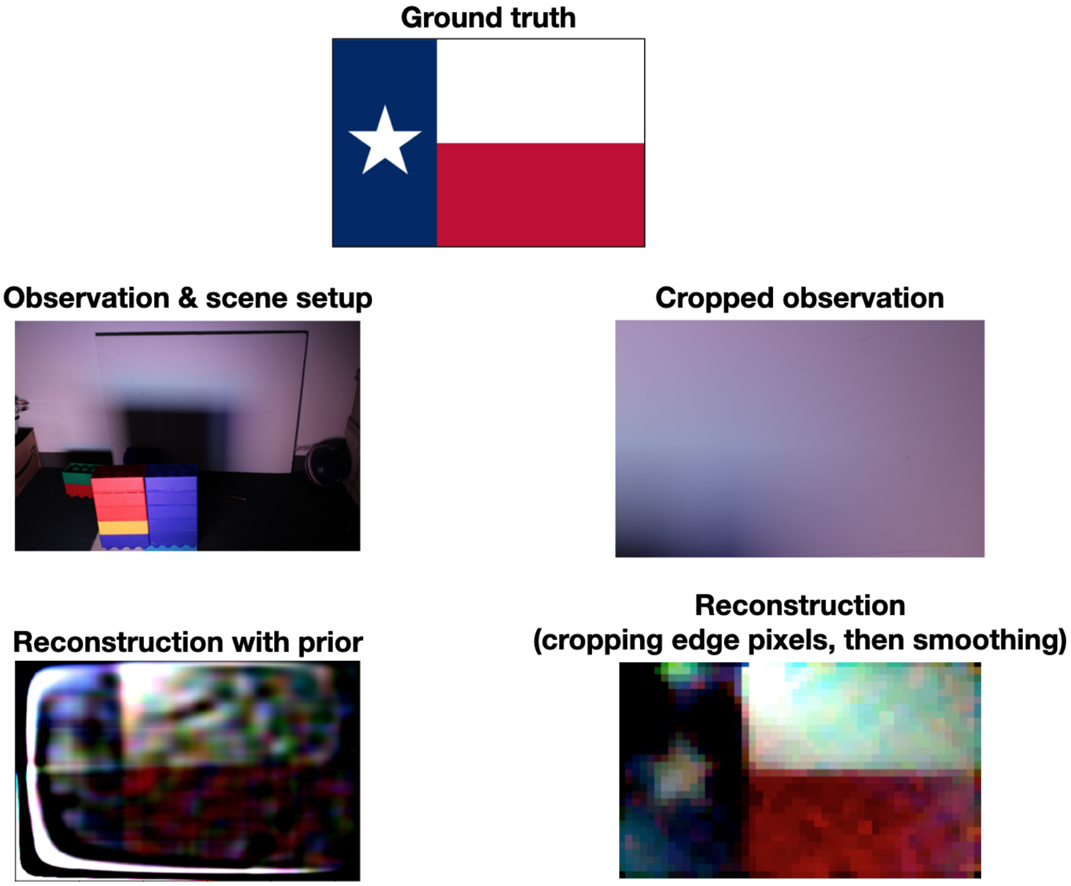
the image! The display you see in 3-40 clips their values from above and below, but if it weren't for that they would be blindingly bright.

This causes serious problems when inverting using a prior, as described in Section 3.6. This is because spatial priors make the assumption that the intensity of the recovered pixels won't vary sharply, but reconstructions that make use of the equivalent occluders in this way rely on recovered pixels varying unrealistically in intensity near the edges of the recovered image. Using a prior and equivalent occluders at the same time can cause nasty artifacts to appear when the priors cause the degenerate pixels to "bleed" into the rest of the image. In such cases, you're better off zeroing out the pixels on the edges, then using a Gaussian blur on the reconstruction, to simulate the effects of a spatial prior. See Figure 3-41.

Many other families of equivalent occluder frames are easily found by finding pairs of occluder frames M_1 and M_2 such that $\frac{d}{dx}\frac{d}{dy}M_1 = \frac{d}{dx}\frac{d}{dy}M_2$. Figure 3-42 gives an example of one.

3.20 Non-parallel Occluder Imaging

In Section 4.1, a method for imaging scenes around a corner is described. For the most part (ignoring the stereo camera), the only reconstructions we get are 1-dimensional, because we integrate along one of the dimensions of space on the observation plane. This method is closely analogous to the method we might use assuming the standard configuration, with an occluder that was half-on and half-off (see Fig. 3-43). Yet, despite being closely analogous, the two situations are not the same; in the former case, we sum the intensities radially and differentiate along the angular dimension to reconstruct a 1D view of the scene, and in the latter case, we sum the intensities along one Cartesian dimension and differentiate along the other in order to reconstruct the scene. This difference is owed to the fact that in the former case, the occluder is perpendicular to the observation plane, whereas in the latter case, it's parallel; but how can we generalize this? How will our reconstruction algorithms behave as a function of the angle of a planar occluder to the scene?



Ground truth

Observation & scene setup

Cropped observation

Reconstruction with prior

Reconstruction
(cropping edge pixels, then smoothing)

Figure 3-41: An experiment with a right-angle occluder (the stack of legos) where one quadrant occludes and the other three do not. This occluder is not separable, but it is equivalent to one that is. The ground truth image was displayed on a monitor to yield the observation. Thus the reconstruction can be obtained simply by taking the derivative of the cropped observation along both spatial dimensions. Note that when using a prior, this causes large, ugly artifacts to appear, especially near the edge of the image.

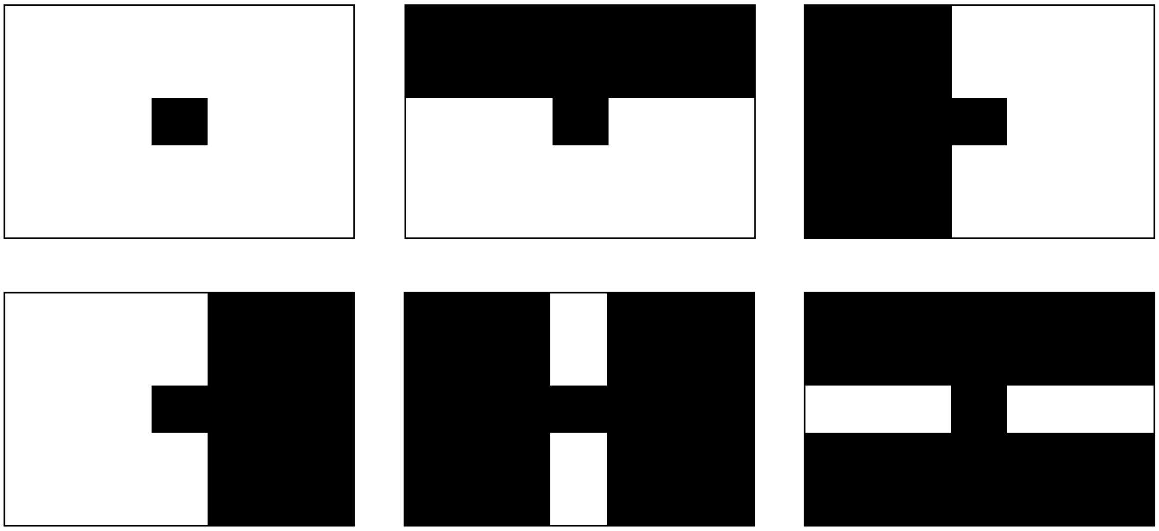


Figure 3-42: Another family of equivalent occluders. Any of these occluders can be freely substituted for any of the others, and most of the pixels of the resulting image will be exactly correct.

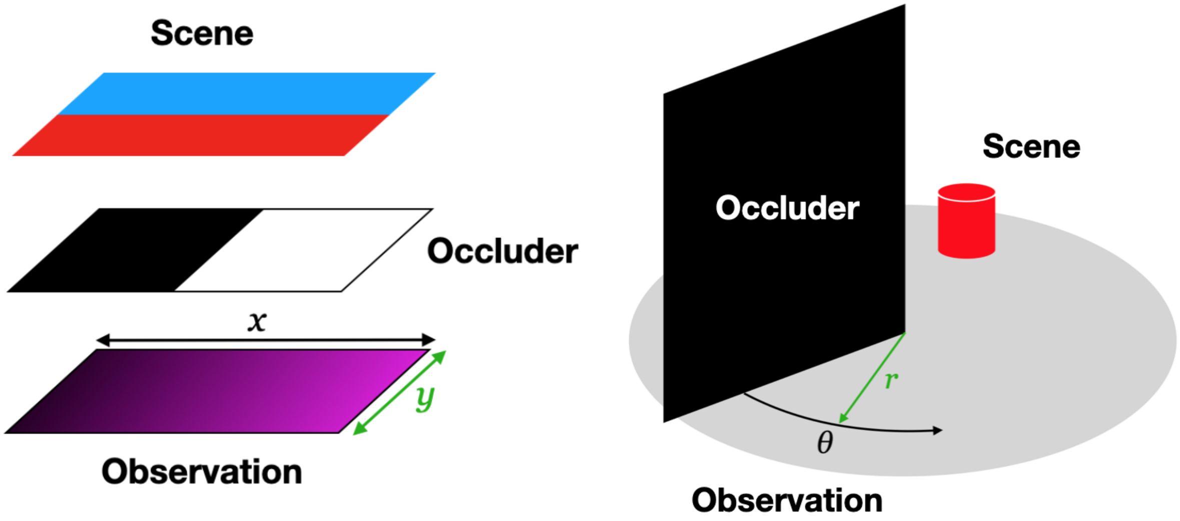


Figure 3-43: Left: a configuration with an edge occluder parallel to the scene. To get a 1D reconstruction of the scene, we sum the intensities along the y dimension and differentiate along the x dimension. Right: a configuration with an edge occluder perpendicular to the scene. To get a 1D reconstruction of the scene, we sum the intensities along the r dimension and differentiate along the θ dimension.

Sections 3.11 and 3.16 describe how to handle occluders at weird locations, or non-planar occluders, so we could simply voxelize the space in between scene and observation, and simulate a plane that's at a weird angle as several small occluders at different depths. This is unsatisfying, though, and causes us to miss out on some fascinating geometry that will more cleanly describe the phenomenon we see in Figure 3-43.

At this point, if you can, I recommend getting out a sheet of paper (or any other opaque flat object made up of straight lines—a book will do as well) and hold it up, at a downward-slanted angle, above a flat surface such as a table, using your left hand. If you're using a sheet of paper, do your best to hold it rigid; don't let it droop. With your right hand, hold a flashlight, pointed downward (such as the one on your smartphone). Keeping the paper in place, move the flashlight around—horizontally, vertically, whatever you like. You should see two lines on the table that define the shape of the paper's shadow; those two lines will meet at the paper's corner's shadow. Let's call those lines u and v, respectively. Of course, those two lines will move as you move your flashlight around. So really, there's some set U of possible lines u that will appear on the table, depending on where you're holding the flashlight; similarly, there's a second set V of possible lines v.

At this point, you may notice that all of the lines u in U all meet at a single point, and that the same is true of all of the lines v in V. This is remarkable; a set of more than two lines won't necessarily have a single point where they all meet. Let's call these two points p_u and p_v at which all lines meet *focal points*. The two focal points of your occluder are the points at which, if you continued the two lines that form the edges of your sheet of paper downward, those lines would intersect the table. If any of this paragraph or the previous one didn't make sense, see Figure 3-44, which demonstrates the proper technique for holding occluders above tables, and shows where p_u and p_v are on the table given the angle of the occluder.

These two focal points tell us how to reconstruct from a given occluder made up of straight lines, suspended above the observation plane at some arbitrary angle. We have to find the focal points of those lines, and it will be our angle from each of those

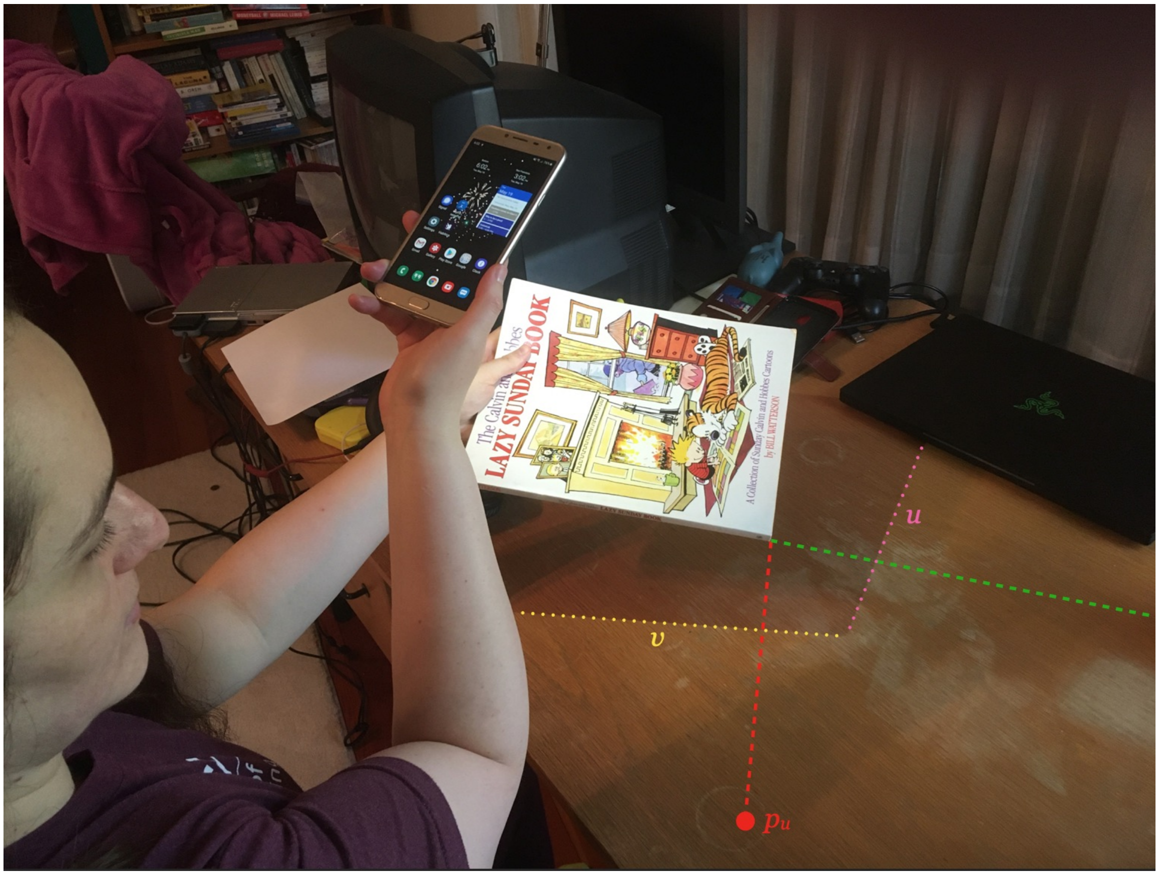


Figure 3-44: A woman holding a flashlight above a book with one hand, and holding a book above the table with the other, as described in the third paragraph of Section 3.20. If she moves the flashlight around, it will move the shadow edges u and v, but u and v, if extended, will always go through p_u and p_v, the focal points of this occluder, no matter where the flashlight may be. The focal points p_u and p_v can be found by extending the edges of the book downward until they intersect the table. p_v (in green) is to the right of the frame; it would appear when the green dotted line intersected the table.

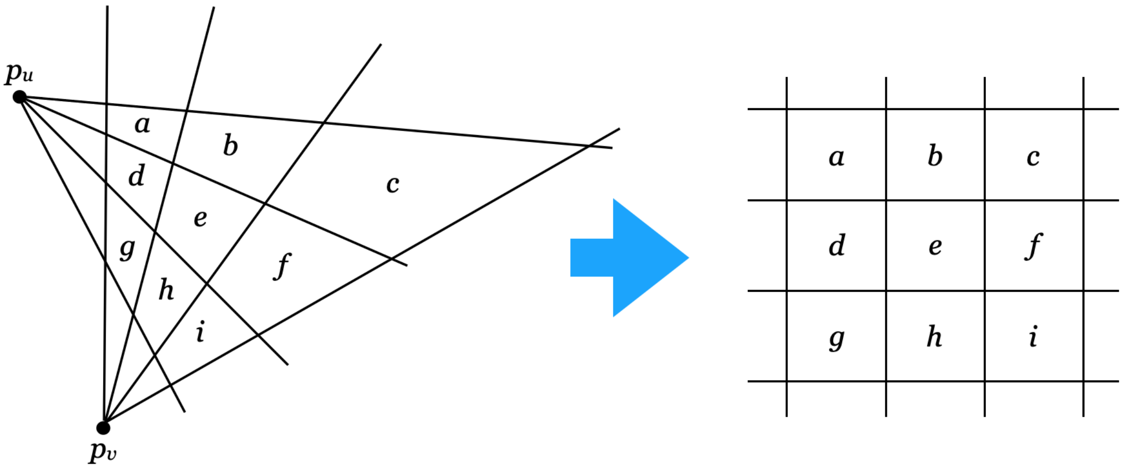


Figure 3-45: The transformation from locations on the observation plane to points in the scene, for a given pair of focal points.

focal points that tells us where points in the scene might be. See Figure 3-45 for an example transformation from locations on the observation plane to locations on the scene, for a given pair of focal points.

And now we can understand why, when the occluder is perpendicular to the floor, we have a focal point at the point where the occluder meets the floor. And when the occluder is *parallel* to the floor, well, the two focal points are at infinity! No transformation is necessary then, since the angle from a point at infinity is equivalent to Cartesian distance.

Given a right-angle occluder such as a book or a piece of paper, we can reconstruct the scene above by performing the above-described transformation, then differentiating along both Cartesian coordinates in the transformed space (the right space shown in Fig. 3-45). It should even be possible, in principle, to blindly reconstruct the angle from the observation plane of a right-angle occluder, by measuring the directions of greatest variation on the observation plane (those directions are likely to be perpendicular to the focal points).

3.21 Optimal lenses under thickness and curvature constraints

This section examines what happens when we explore not only occluding aperture frames but aperture frames that can redirect the light as well. Being able to freely redirect light corresponds to the constraint of considering only left-stochastic matrices (i.e. matrices whose columns must sum to 1). This is because each column of the transfer matrix corresponds to the impulse response to the light in one location. If each column of the transfer matrix sums to 1, that corresponds to saying: all of the light that enters the aperture frame must leave, no more and no less.

Optimizing for mutual information when we can freely redirect the light leads to the natural conclusion that the identity transfer matrix is optimal; it's well-known that over left-stochastic matrices, the determinant is maximized by the identity matrix (or any other permutation matrix) [45]. This transfer matrix can be achieved with a traditional lens, with the scene in the lens' focal plane.

If we limit the thickness of the aperture frame, then a Fresnel lens will be optimal [84] (see Figure 3-46). If we also impose the constraint that the glass cannot be curved, however, we are left with an interesting problem. This has the additional benefit of equalizing the number of degrees of freedom between our study of occluders and our study of glass aperture frames (if we divide our frame into n patches, we are left with 2^n possibilities in both the former and latter case, but we would be left with more possibilities than that in the latter case if we also needed to consider glass curvature). This constraint also potentially reduces the cost of the frame.

This pair of constraints corresponds to a limitation on the distance by which we can redirect light, and the amount we can change the direction of the light. As we can see from Figure 3-47, a flat glass plane will redirect the light increasingly with the incident angle, the thickness of the glass, and the ratio of the refractive index of air n_1 to the refractive index of the glass n_2. Suppose that the scene and observation plane are represented by vectors of length n, and suppose they are separated by a distance y and each have length x. Consider a binary array of aperture frame elements g_i

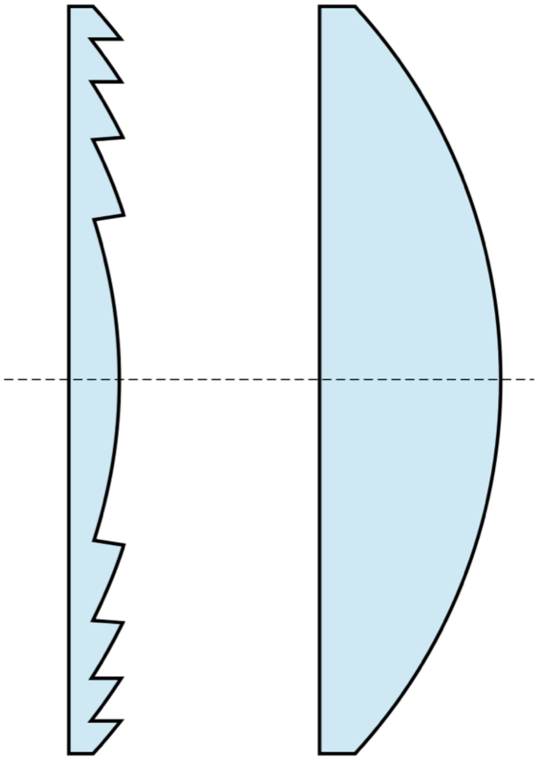


Figure 3-46: Left: A Fresnel lens. Right: A convex lens of equivalent power. Note that the thickness of the glass at each point on the Fresnel lens is equal to that of the convex lens, modulo the maximum thickness of the Fresnel lens. Image credit due to Wikipedia.

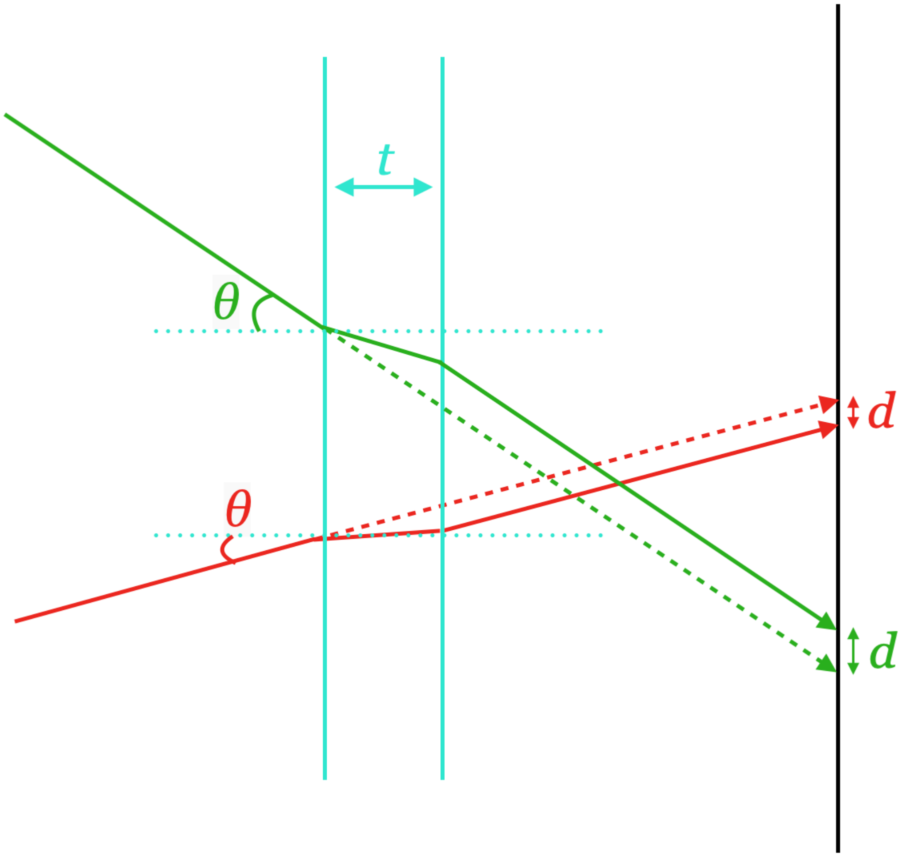


Figure 3-47: The displacement d of an incoming light ray increases with the thickness t of the glass, and with the incident angle θ.

where there is a piece of transparent material at position $x/(2n-1)$ if $g_i = 1$, and no glass-like material otherwise. Assuming there is glass-like material, suppose it has thickness t and refractive index n_2 (and we'll take the refractive index of air to be 1). In that case, we can use Snell's law to build the transfer matrix corresponding to the flat glass array g_i as follows:

$$A \leftarrow \mathbf{0}$$

Initialize the transfer matrix to the all-zeros matrix.

$$\forall 0 \leqslant i, j < n$$

$$\Delta x = \frac{x}{2n}(i - j)$$

This Δx is the light's displacement on its way from the scene to the aperture frame (this is ignoring the thickness of the aperture frame, assuming $y \gg t$).

$$l = \frac{y}{n(y^2 + \Delta x^2)}$$

The amount of light that would hit the observation from the point X_i in the scene to Y_i on the observation if the glass were absent, accounting for near-field effects.

$$\theta_i = \tan^{-1}\left(\frac{2\Delta x}{y}\right)$$

The angle of the incident light to the aperture frame, relative to the aperture frame's normal.

$$\theta_g = \sin^{-1}\left(\frac{n_1}{n_2}\theta_i\right)$$

The angle of the light while it's inside the glass, derived using Snell's law.

$$d = \frac{2n}{x}tg_{i+j}(\sin(\theta_g) - \sin(\theta_i))$$

The displacement d caused by the glass.

$$A_{i+d,j} \leftarrow A_{i+d,j} + l$$

Increment the corresponding element of A by the amount l. Note that while using finite matrices A, d may not be an integer. If this is the case, we can use a linear approximation by incrementing each of $A_{\lfloor i+d \rfloor,j}$ and $A_{\lceil i+d \rceil,j}$ by amounts scaled by $1 - rem(d, 1)$ and $rem(d, 1)$, respectively.

Using this procedure, we can compute the transfer matrix A for a given glass array g. Figure 3-48 shows the performance of optimal glass frames for three different glass thicknesses and three different refractive indices, found using exhaustive search.

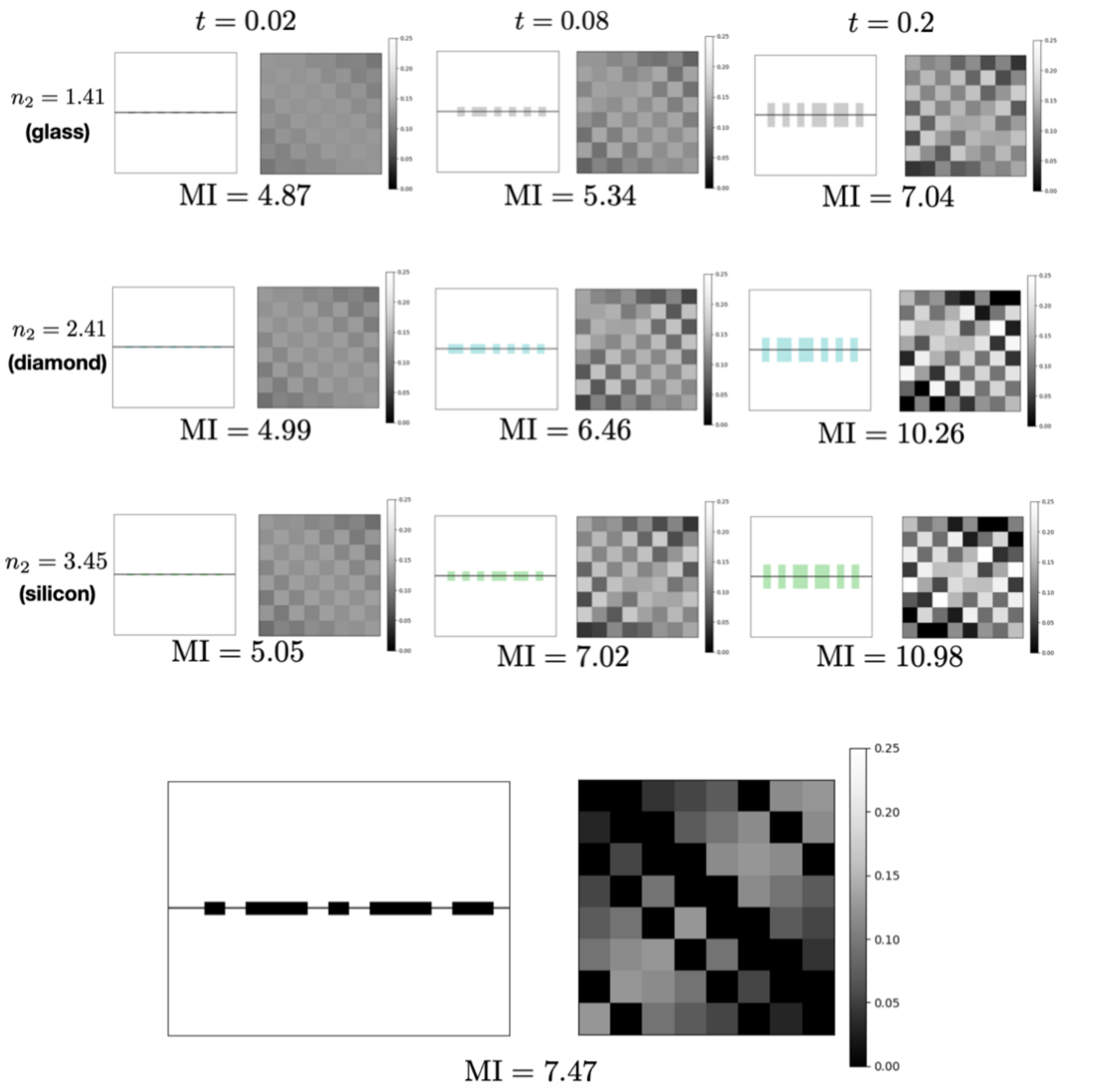


Figure 3-48: Top: the performances of optimal transmissive aperture frames for three different choices of n_2 (corresponding to the approximate refractive index of glass, diamond, and silicon, respectively), and for three different choices of the glass's thickness. Unsurprisingly, optimal performance improves with higher thickness and higher refractive index. Other parameter choices are $n = 8$, $x = 1$, $y = 1$, the SNR $\sigma = 10^3$, and the scene correlation coefficient $\beta = 0.1$. Each sub-image shows the form of the aperture frame on the left, and the corresponding transfer matrix on the right. Bottom: for comparison, the performance of the optimal occluding aperture frame, under identical conditions. As can be seen, using an occluder substantially outperforms all but the thickest glass, if the glass is assumed to be flat.

3.22 Optimal phase arrays

The earlier bound in Equation 3.8, implied by Hadamard's bound, on the determinants of circulant matrices assumed binary (meaning $\{0, 1\}$) matrices. To recall, that bound was:

$$|\det B| \leqslant 2^{-n}(n+1)^{(n+1)/2}$$

for a binary matrix B. For $\{1, -1\}$ matrices C, recall that Hadamard's bound directly states:

$$|\det C| \leqslant n^{n/2}.$$

What about for matrices D that contain any complex number with norm at most 1? In a real-world setting, this might correspond to an aperture frame capable of modulating the phase of the incoming light. In this case, the optimal sequence d_i (i.e. the first row of D) is a complex quadratic, and is given by:

$$d_k = \exp\left(\frac{2\pi i(k^2 - k)}{2n}\right), k \text{ odd}$$

$$d_k = \exp\left(\frac{2\pi i(k^2 - 2k + 1)}{2n}\right), k \text{ even}$$

Note that the pattern of a quadratic with a modulus applied closely mirrors the Fresnel lens (see Figure 3-46).

Chapter 4

Occluder-based Non-line-of-sight Imaging

4.1 Turning Corners into Cameras

4.1.1 Introduction

Most of the content in this section is taken from *Turning Corners into Cameras: Principles and Methods* [16]. My contributions to this project were mainly in helping to run many of the experiments, making figures, editing the writing, and analysis of the effects of corner errors. This publication describes an occluder-based imaging system that maps neatly onto the "standard configuration," using an "edge occluder" as the occluding aperture frame. An "edge occluder" means an occluding frame that lets through half the light and occludes the other half. This section will summarize the methods and results of [16], and provide some analysis of edge occluders within the framework presented in Chapter 3.

This particular type of occluder also has the additional nice property of having, like the pinhole, a particularly recognizable reconstruction algorithm. A pinhole inverts (and blurs) the scene but otherwise doesn't distort it, so that the observation matches the image; an edge occluder integrates the scene, so that if the scene intensity function is $f(x)$, the observation intensity function will be $\int f(x)dx$. That means that in

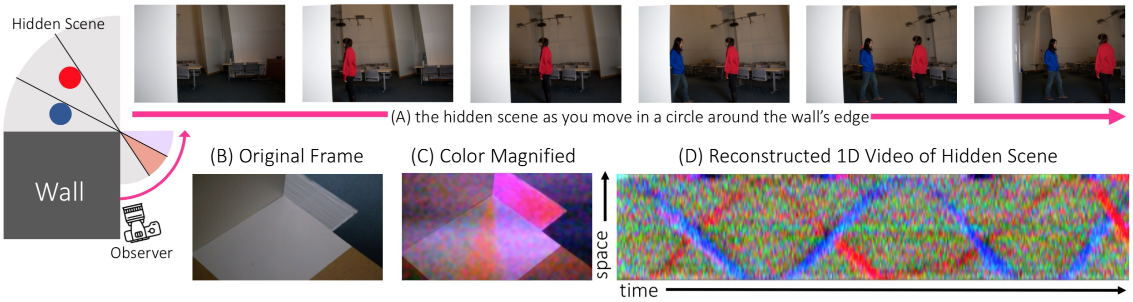

Hidden Scene
Wall
Observer
(A) the hidden scene as you move in a circle around the wall's edge
(B) Original Frame
(C) Color Magnified
(D) Reconstructed 1D Video of Hidden Scene
space
time

Figure 4-1: A method for constructing a 1-D video of an obscured scene. The far left shows a diagram of a typical scenario: two people—one wearing red and the other blue—are hidden from view by a wall. To an observer walking around the occluding edge (along the magenta arrow), light from different parts of the hidden scene becomes visible at different angles (A). Ultimately, this scene information is captured in the intensity and color of light reflected from the corresponding patch of ground near the corner. Although these subtle irradiance variations are invisible to the naked eye (B), they can be extracted and interpreted from a camera position from which the entire obscured scene is hidden from view. Image (C) visualizes these subtle variations in the highlighted corner region. We use temporal frames of these radiance variations on the ground to construct a 1-D video of motion evolution in the hidden scene. Specifically, (D) shows the trajectories over time of hidden red and blue subjects illuminated by a diffuse light in an otherwise dark room. Figure originally from [16].

order to reconstruct, all you have to do is take the derivative of your scene with respect to space. This is convenient not only because it means the core algorithm is simple, but also because some cameras already exist which automatically record spatial derivatives at the hardware level [73].

The headline figure from *Turning Corners into Cameras* explains this idea well; see Fig. 4-1. Note how the colors on the floor give us a 1D view of what's in the scene, integrated over space. This is exactly what we expected to see from an edge occluder, given our analysis in the previous chapter. The fact that in this case, the observation plane (i.e. the floor) is perpendicular to the occluder frame (i.e. the wall) is what makes the spatial variation on the floor be a function of angle from the wall, rather than just the x-coordinate. This is a direct consequence of the phenomenon discussed in Section 3.20.

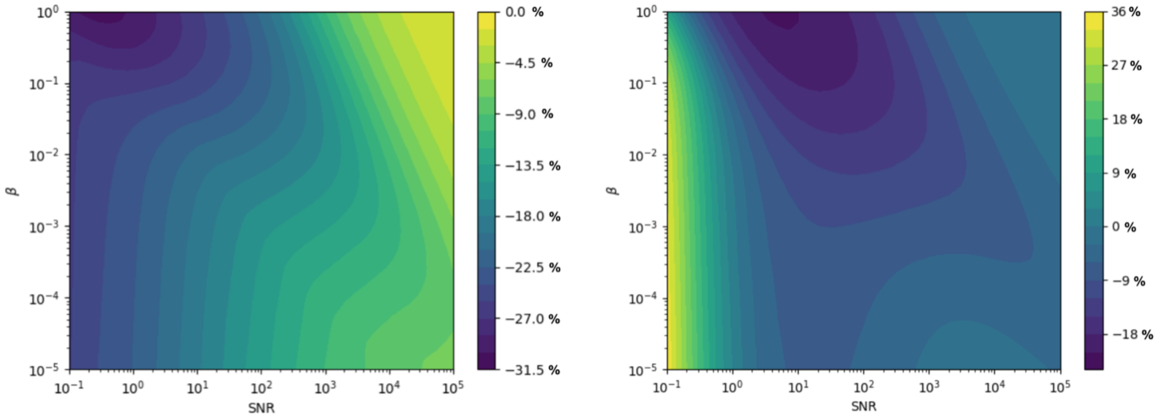


Figure 4-2: Left: the percent difference in performance between the edge camera and *optimal* size of circulant pinhole for each value of SNR and β. Negative values indicate that the pinhole outperforms the edge camera. Right: the percent difference in performance between the edge camera and a circulant pinhole camera with transmissivity $\rho = 1/4$.

4.1.2 A comparison of the edge camera to the pinhole camera

How does an edge occluder compare to other types of 1D occluders? One salient point of comparison is the pinhole. Obviously, the quality of pinholes varies as a function of their size (see Section 3.13), which makes a direct comparison between the edge camera and the pinhole camera challenging. Using Figure 3-28 as a reference for what the optimal pinhole size is at each point in the SNR/scene correlation matrix, we can compare the performance of the edge camera at each such point to whatever the optimal 1D circulant pinhole is at that point. See Figure 4-2 to see this comparison. As can be seen from Figure 4-2, the edge camera is always outperformed by the ideally-sized pinhole, sometimes substantially so; however, perhaps a fairer comparison would be between the edge camera and a single, fixed size of pinhole, since the edge camera has no parameter which can be optimized in the way the pinhole does. We can see that the edge camera performs similarly overall to a pinhole camera with transmissivity $\rho = 1/4$.

Figure 4-2 shows that the edge camera is similar in quality to a 1D pinhole camera. We know, however, that pinhole cameras are far from optimal. The thing that makes edge cameras interesting is not the direct interpretability of the observation plane (as it is for a pinhole camera) nor is it optimality or near-optimality (as for a spectrally flat occluder). Rather, it's their ubiquity. Anywhere that a wall corner blocks your vision, an edge camera is available for use in reconstructions. And the fact that this method is relatively less ambitious, yielding only a 1D reconstruction, means that the SNR of the method will be correspondingly much higher, because we can average over an entire dimension.

4.1.3 Averaging in space and in time

The impressive robustness and practicality of this method is due to the trick of averaging in space and in time to reduce the variance and the bias of our reconstruction, respectively. Briefly, I will explain how and why this works in the language of the previous chapter.

First, we can improve the effective SNR of the imaging system—thereby reducing the variance of the reconstruction—by averaging over the dimension parallel to the edge, in the simplified setting where scene, observation, and occluder are all parallel. Assuming the paraxial approximation, any reconstruction would be uniform along that dimension anyway, given the form of the occluder; there is no way to distinguish the location of a light source in the scene along that dimension.

In the real-world setting where the occluder lies perpendicular to the observation, this "dimension of constancy" will no longer be parallel to the edge, but will rather extend outward radially from that edge (see Section 3.20 for more discussion). This means that the spatial averaging will take place instead over a narrow triangular slice of the observation. See Figure 4-3 for an illustration.

Second, we can correct for the effect of varying albedo on the observation, as well as the effects of nuisance light illuminating the observation non-uniformly by subtracting the average of our observation over time. This will reduce the bias of our reconstruction. Recall that we intend to reconstruct the scene from the observation basically by taking a spatial derivative of what we see on the observation; but if our observation plane is a non-uniform patterned floor, the subtle effects we hope to measure from the light coming from the scene will be drowned out by the patterns on

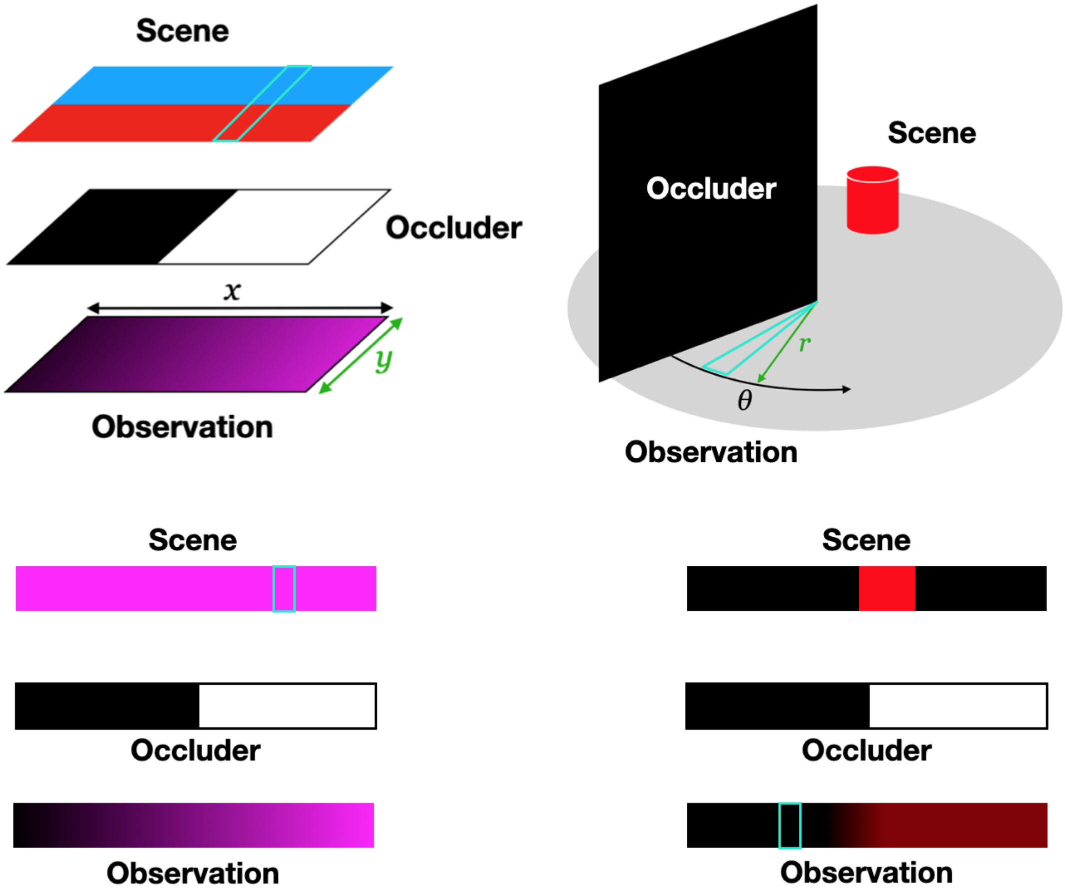


Figure 4-3: Top left: an idealized setting in which scene, observation, and occluder are all parallel. Bottom left: a simplified version of the setting in flatland, in which the region highlighted in cyan in the top left is averaged to produce the region highlighted in cyan in the bottom left. Top right: the real-world setting in which the occluder lies perpendicular to the observation. Bottom right: a simplified version of the setting in flatland, in which the region highlighted in cyan in the top right is averaged to produce the region highlighted in cyan in the bottom right.

the floor. We would like to take a spatial derivative not of our raw observation, but of our observation after subtracting away the floor's pattern. Ideally, we would know the floor's pattern by observing what it looks like in response to an "empty scene"; in practice, we usually have no way to know that, since we don't know what's in the scene, so we don't know when it's empty.

So as a proxy for what the floor's pattern is, we can instead use the average over a long time period of what the floor looks like. We hope that after averaging over that time period, the resulting time-averaged scene is approximately uniform. This lets us correct for nuisance non-uniformities in the imaging system.

Unlike the previous idea, this idea is very general, and tremendously useful in uncalibrated passive non-line-of-sight imaging systems; it will get used again in Section 4.5.

4.1.4 Other differences

Of course, the reconstruction algorithm isn't purely a spatial derivative; that's an oversimplification. In reality, the method of [16] relies on a maximum a posteriori (MAP) estimate of the scene using an L2 regularizer. While not exactly the same as the reconstruction algorithm described in Section 3.6, it is similar. From Figure 4-4, we can see that the reconstruciton algorithm roughly corresponds to a "blurry" spatial derivative.

4.1.5 Selected Results

In the following sections I'll show a few results from the original paper.

The algorithm reconstructs a 1-D video of a hidden scene from behind an occluding edge, allowing users to track the motions of obscured, moving objects. In all results shown, the subject was not visible to an observer at the camera.

Fig. 4-5 shows a few examples of 1-D videos recovered from indoor edge cameras. In these sequences, the environment was well-lit. The subjects occluded the bright ambient light, resulting in the reconstruction's dark trajectory. Note that in all the reconstructions, it is possible to count the number of people in the hidden scene, and to recover important information such as their angular size and speed, and the characteristics of their motion.

Outdoor: In Fig. 4-6 we show the results of a number of videos taken at a common outdoor location, but in different weather conditions. The top sequences were recorded during a sunny day, while the bottom two sequences were recorded while it was cloudy. Additionally, in the bottom sequence, raindrops appeared on the ground *during* recording, while in the middle sequence the ground was fully saturated with water. Although the raindrops cause artifacts in the reconstructed space-time images, you can still discern the trajectory of people hidden behind the wall.

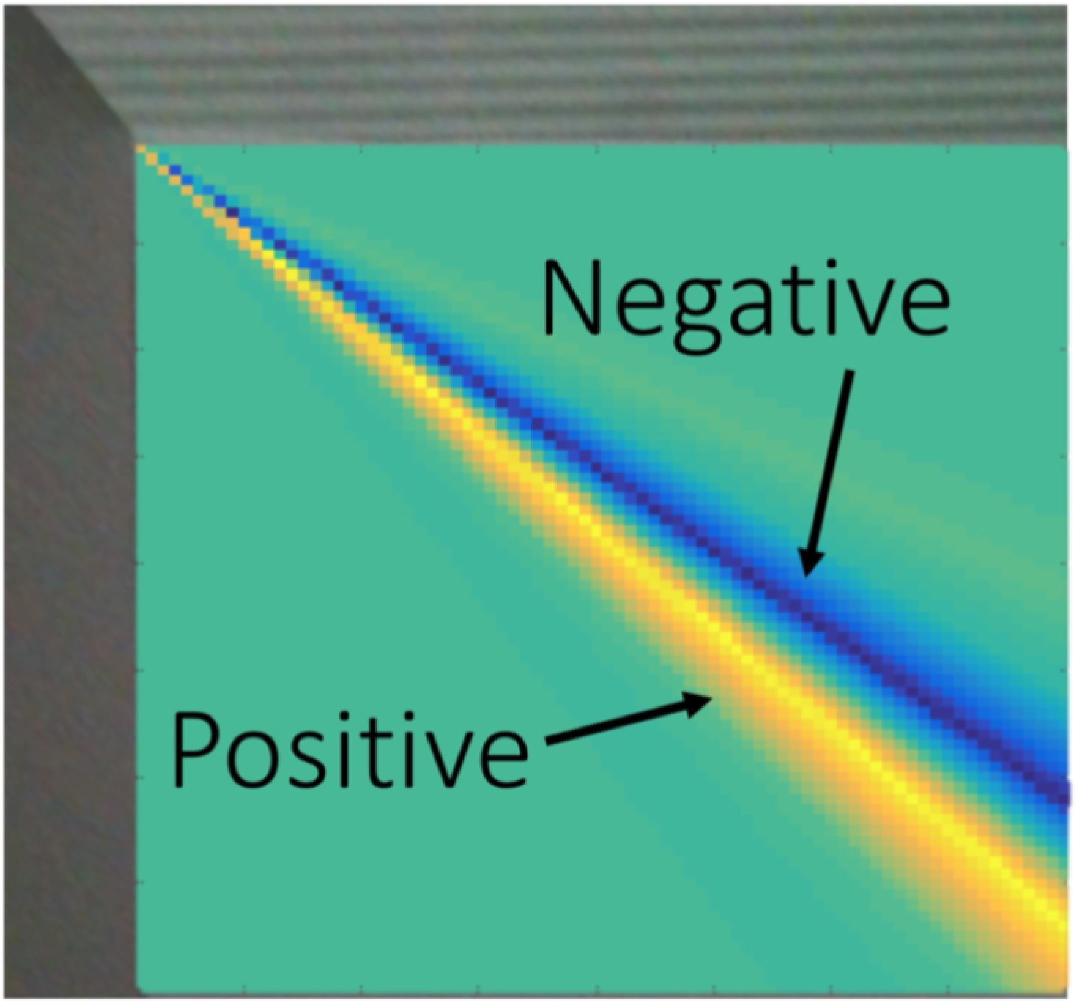


Figure 4-4: The example estimation gain image, showing the operation performed on the observation to recover the scene. As we can see, it's approximately a spatial derivative along the angular dimension. The spatial derivative being taken has a "blurry" appearance because of the spatial prior. Figure originally from [16].

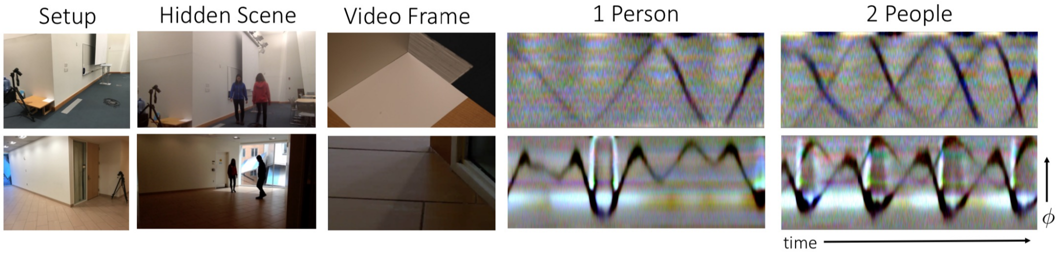


Figure 4-5: Indoor experiments for the corner camera. The subjects' trajectories are clear from the traces in the reconstructed 1D movie (right). Note that the x-axis corresponds to time. Note, also, that the number of subjects can be easily counted.Figure originally from [16].

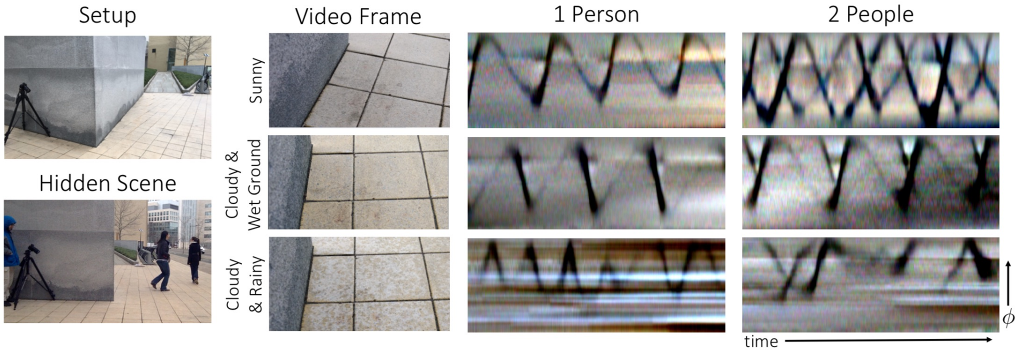


Figure 4-6: Outdoor experiments for the corner camera. In sunny weather, the subjects' trajectories are very clear. In cloudy weather, the trajectories are fainter but still clear upon closer examination. In rainy weather, the raindrops create dark streaks in the reconstructed movie trace. However, if we ignore those streaks, we can still see most of the subjects' trajectories.Figure originally from [16].

See the paper itself [16] details about the method and for more results.

4.2 The effects of corner errors

One important source of error in the edge camera is the *corner location error*. When studying a movie of the projection plane, it's important to know where the corner of the wall is in order to make an accurate reconstruction. Corner location errors occur when the corner of the wall is erroneously chosen to be the wrong place. Corner location errors introduce systematic error into the scene's reconstruction.

The corner occluder can be chosen automatically (by minimizing the intensity variation along the “rays” emanating from the corner onto the projection plane) or manually (by clicking by hand on where the corner is in the image). Both of these methods have the potential to introduce a corner location error.

4.2.1 Edge Camera

Exactly how bad are corner location errors? To answer this question, we consider the situation shown in Fig. 4-7. Imagine a dark scene with a single bright object. We want to find the angular position of the bright object in the scene. We can do this by measuring θ: the angle of the shadow it casts against the wall. When we find the angle at which the projection plane goes from light to dark, we will know what θ is.

This story is simple in the case when there is no corner location error. But what about the case where there is such an error?

Fig. 4-8 shows this scenario. We can “sweep” the angle ϕ across the projection plane, and at the point where ϕ is midway between dark and light, we can presume that that is the object’s angular position. When there is no corner location error, we will naturally get $\theta = \phi$, but when there is a corner location error, ϕ will depend on θ and other parameters in a more complicated way.

Fig. 4-9 plots intensity against the sweeping angle ϕ, both with and without a corner location error. Note that in the case where there is a corner location error, the maximum intensity value no longer takes on a maximum value of 1, but a value below 1. In the analysis that follows, we will call that value l_{max}, and we will use $l_{\text{max}}/2$ as the “transition point” between light and dark. In other words, we will choose the ϕ that gives an intensity of $l_{\text{max}}/2$ as our estimate for θ.

Fig. 4-10 is a detailed illustration of the situation, showing the names for the variables that we’ll use in our analysis. As the figure shows, we are presuming a corner location error of (d_x, d_y) and a projection plane radius of r. We want to find what our estimate θ, ϕ, will be as a function of θ and in terms of d_x, d_y, and r.

Using Fig. 4-10 as a reference, we can make the following observations:

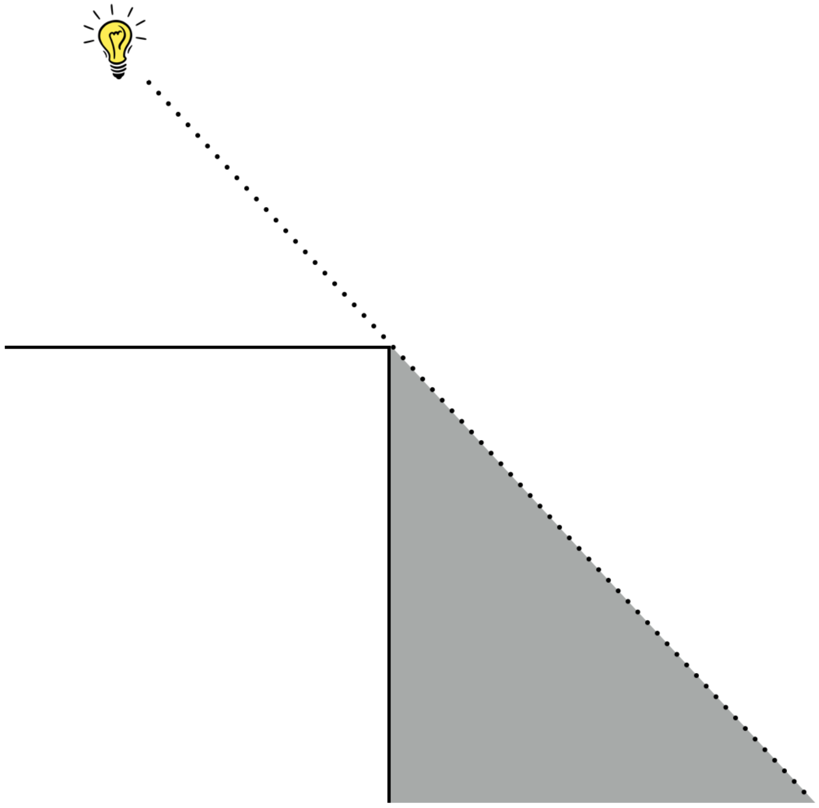


Figure 4-7: This figure shows the configuration for the toy problem of interest. The scene consists of a single bright object, whose angular position θ we want to learn.

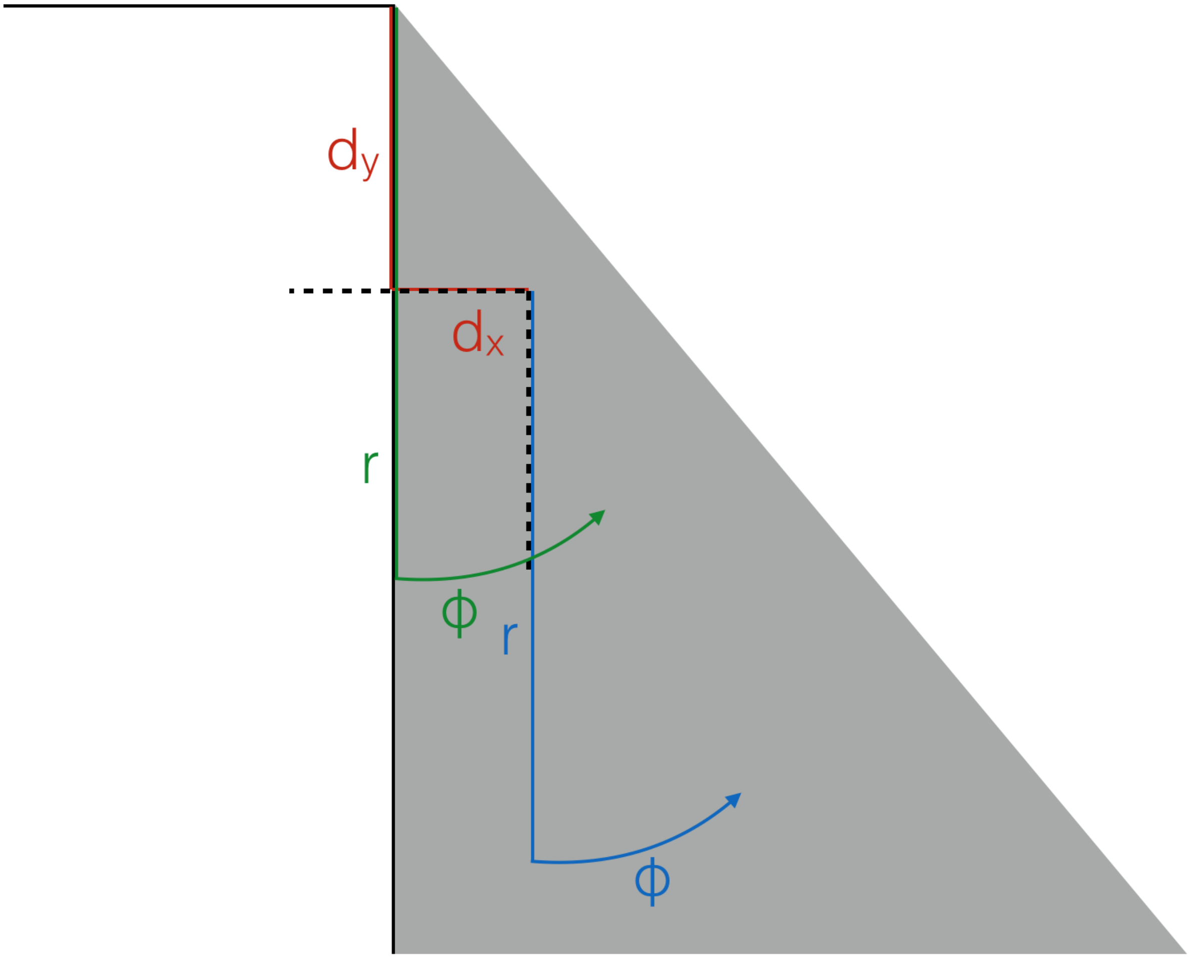


Figure 4-8: This figure shows the impact of a corner location error. In the error-free case, we would sweep ϕ across the projection plane (shown in green) hinging around the corner (the solid black line). But if we made a corner location error, we would instead try to sweep ϕ across the projection plane erroneously (shown in blue) hinging around the false corner (shown with a dotted line).

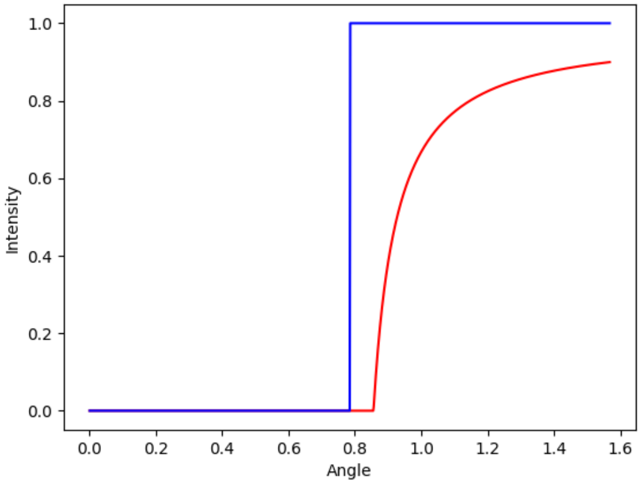


Figure 4-9: This plot shows how the observed intensity values vary with ϕ, in the case of correct corner location (in blue) and a corner location error ($(d_x, d_y) = (0.1, 0.2)$, in red). Note that in the case of a corner location error, the maximum value of the intensity does not reach 1. Note also that d_x and d_y are as a fraction of the radius of the projection plane, r, which here is taken to be 1.

$$\beta = \tan^{-1}\left(\frac{d_x}{d_y}\right)$$

$$l_{\max} = r - d_y \tan(\theta) + d_x$$

$$f = r - \frac{l_{\max}}{2}$$

$$\alpha = \sin^{-1}\left(\frac{\sqrt{d_x^2 + d_y^2}\sin(\theta - \beta)}{f}\right)$$

$$\gamma = \pi - \alpha - \theta + \beta$$

$$\phi = \pi - \gamma + \beta$$

This is how ϕ is expressed in terms of the parameters of the problem (θ, d_x, d_y, r).

What sort of error does this introduce? In order to study this question, we assumed that d_x and d_y were normally distributed with means of 0 and small (relative to r^2) variances σ_x^2 and σ_y^2. We generated many sample (θ, ϕ) pairs for each θ between 0 and $\pi/2$. We then measured the empirical means and variances of these pairs. Fig 4-11

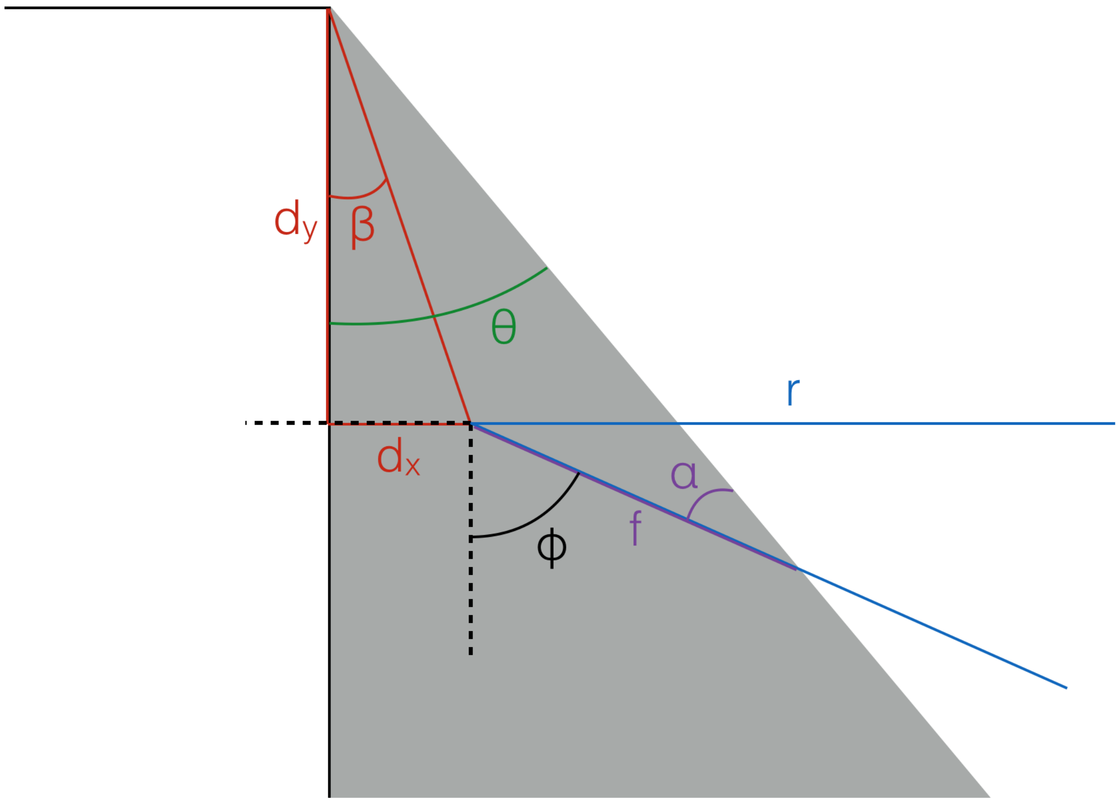


Figure 4-10: This plot is intended as a reference for the meanings of each of the variables used in the calculations of ϕ as a function of θ.

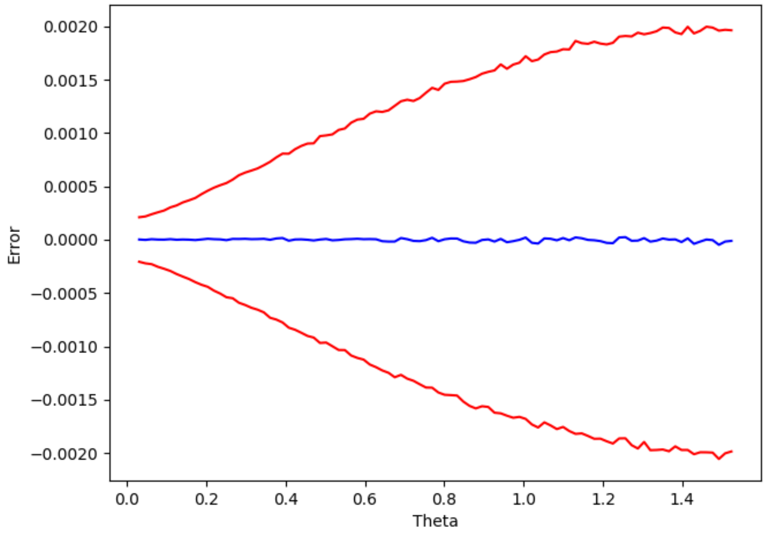


Figure 4-11: This plot shows the empirical mean (in blue) plus or minus one standard deviation (in red) of the error as a function of θ. Here, $\sigma_x = 10^{-4}$ and $\sigma_y = 10^{-3}$.

shows a few of our results.

Here were a few of our empirical findings:

1. The mean error was always 0 for all values of θ, σ_x and σ_y.
2. When $\sigma_x = \sigma_y$, the standard deviation of the error σ_ϵ was $2\sigma_x$ for all values of θ.
3. When $\sigma_x \neq \sigma_y$, the standard deviation of the error σ_ϵ varied between $2\sigma_x$ (for $\theta = 0$) and $2\sigma_y$ (for $\theta = \pi/2$).

4.2.2 Stereo Camera

Another situation in which it makes sense to study corner location errors is in the case where there is a doorway just before the hidden scene, in which case we can use stereo vision to locate a moving object in two dimensions. What effect do corner location errors have on depth estimates, which are generally quite sensitive to noise? To be more precise, suppose that we call the axis along which the doorway lies the

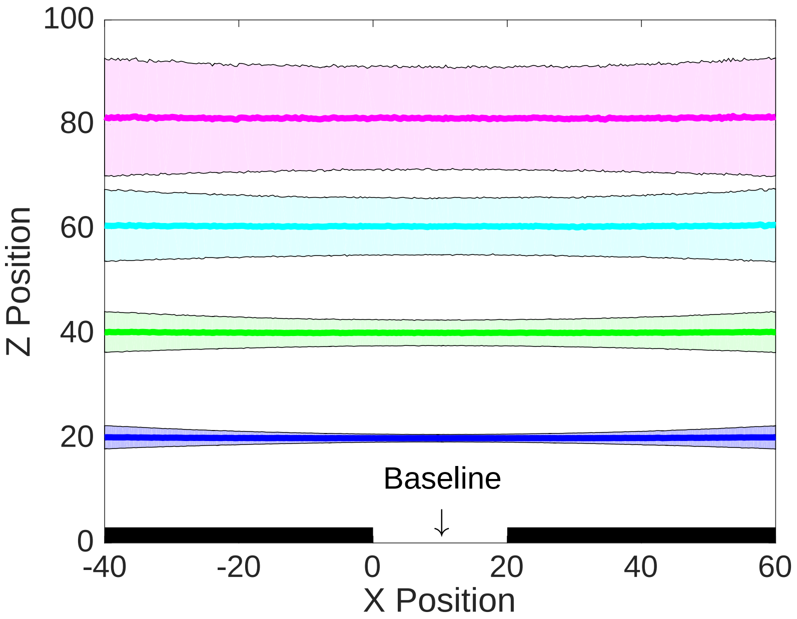


Figure 4-12: The empirical means plus or minus one standard deviation of the estimated P_z as a function of its x-coordinate, assuming true P_z of 20, 40, 60, and 80. Here, the two corner location errors at each of the boundaries of the doorway are independent and subject to $\sigma^2_{\Delta x} = \sigma^2_{\Delta z} = 0.04$. We sample from a set of 1000 corner errors to approximate the mean and standard deviations empirically.

"x-axis," and suppose we call the perpendicular axis (of depth into the room) the "z-axis." Then, how much noise in the z dimension will a corner location error cause?

To give an approximate sense of how much error results in the recovered z position, we show the mean $+/-$ one standard deviation in the z dimension as a function of the true x-position of the object in Fig. 4-12

Note that the empirical means are centered at the true depths of the objects. This does *not* mean that any single corner location error won't cause the depth of the reconstruction to be off systematically, only that on average, corner location errors that are normally distributed around the corners in question will push the reconstructed depths away as much as they pull them closer.

To see this systematic bias on its own, we can also study how a single corner error introduces systematic error in our reconstructions—after all, for a single experiment, we are likely to make a single corner error, and the resulting error in the depth calculations will extend across many x-coordinates as the subject of the experiment

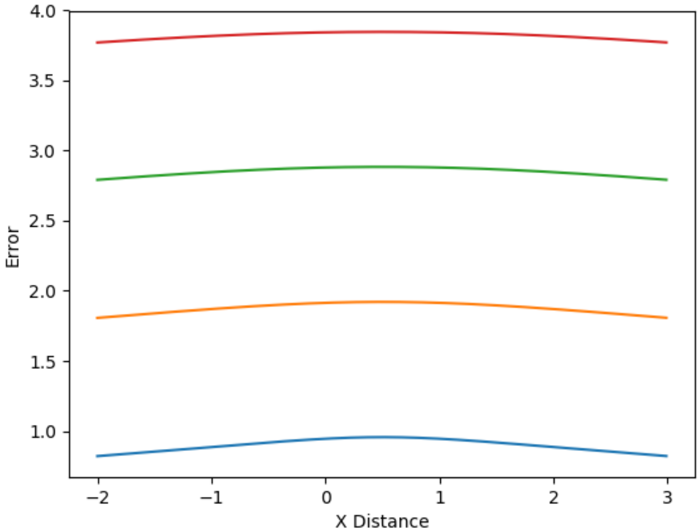


Figure 4-13: The reconstructed depths of objects at depths 1, 2, 3, and 4, given a corner error of $\Delta y_1 = \Delta y_2 = 0.02$.

walks back and forth in the hidden scene. Figs. 4-13 and 4-14 show the systematic bias for two distinct *specific* corner location errors.

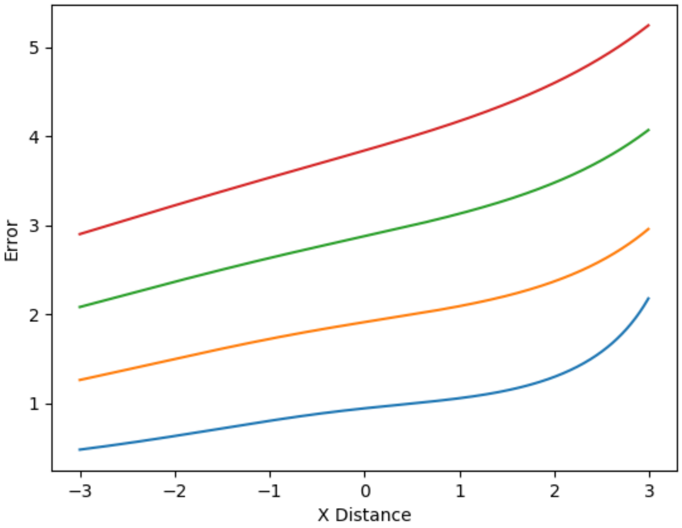


Figure 4-14: The reconstructed depths of objects at depths 1, 2, 3, and 4, given a corner error of $\Delta y_1 = -0.02$, $\Delta y_2 = 0.02$. Note that because of the different corner errors for each corner, there is the possibility of asymmetric behavior on either side of the doorway.

4.3 Inferring Light Fields from Shadows

The content in this section is taken from *Inferring Light Fields from Shadows* [12]. Having joined late in the project, my main contributions to this work were the writing and figures, as well as a few of the later experiments. This section will briefly summarize the methods and results of [12], providing a reader unfamiliar with the topic with an introduction to occluder-based light-field reconstruction.

The goal of the system described in [12] is straightforward but ambitious: using an known occluder, can a 4D light-field be reconstructed from a 2D passive NLoS observation? The occluder is *not* assumed to be planar, nor is the scene. The task here is very difficult, as it is ill-posed: it requires reconstructing a 4D object from a 2D observation (and the 4D known occluder).

Key to this method is the strategy of estimating the transfer matrix with precalibration, as shown in Figure 4-15. Note that this method of calibration is directly analogous to the very definition of the transfer matrix presented in Chapter 1 (see Figure 1-9)!

4.3.1 Overview

In order to reconstruct light fields using secondary reflections from the scene, our imaging method has two main components. The first is a linear forward model that computes observations from light fields, i.e. a transfer matrix **A**, that has many columns but is sparse. The transfer matrix for an arbitrary scene is depicted schematically in Fig. 4-16.

The second is a prior distribution on light fields that allows reducing the effective dimensionality of the inverse problem, turning this ill-posed problem into one that is well posed and computationally feasible. This strategy is better than other methods for reducing the dimensionality of the inverse problem (for example, naively downsampling the forward model and inverting). Light field sampling theory [24] and novel light field priors [65] inform how we reduce the dimensionality of the inverse problem given mild assumptions of the elements that produce the light field to be recovered.

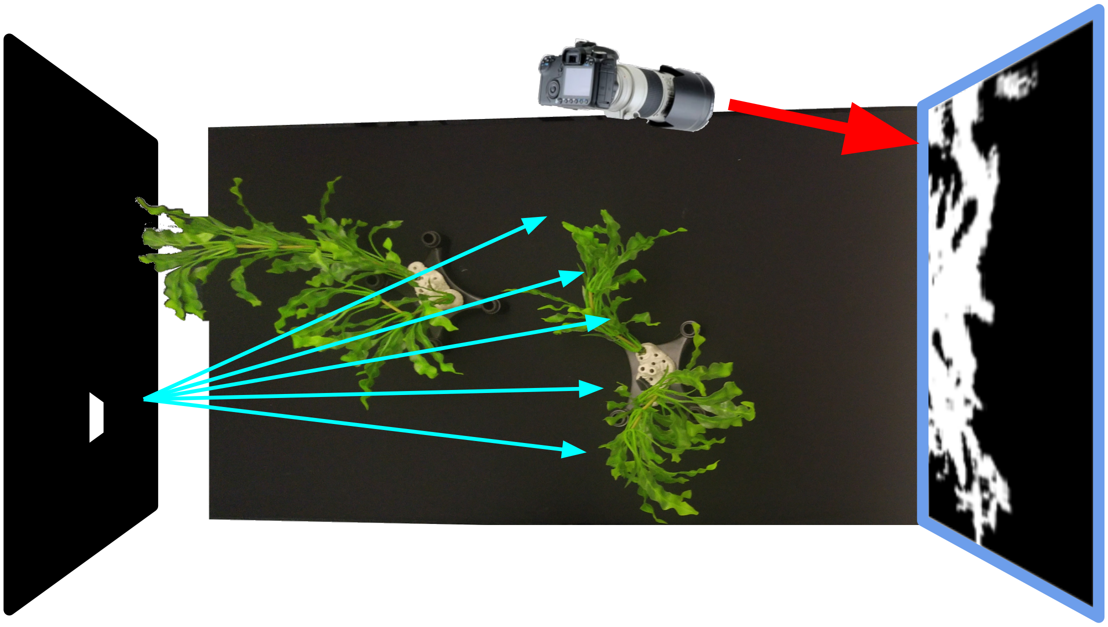


Figure 4-15: The process of calibrating for the occluder, performed by lighting one small region of the monitor on the left at a time, and recording the shadow on the right that results. Note the similarity of this process to the definition of the transfer matrix from Chapter 1.

The work of [12] parametrizes the occluded light-field (meaning the light-field that results from the presence of both the scene and the occluded) as the elementwise product of the native scene light-field and the occluder visibility function (see Figure 4-17). Note the slanted pattern of occlusion created by the occluders at different depths, as discussed in Sections 3.11 and 3.16.

The details of the method can be found in the paper itself [12].

4.4 Selected Results

This section contains a few results taken from [12]; these results were obtained under heavy illumination. See Figs. 4-18 and 4-19. As we can see, the reconstructions are blurry but recognizable, and the reconstructed light-field includes a noticeable and realistic parallax effect.

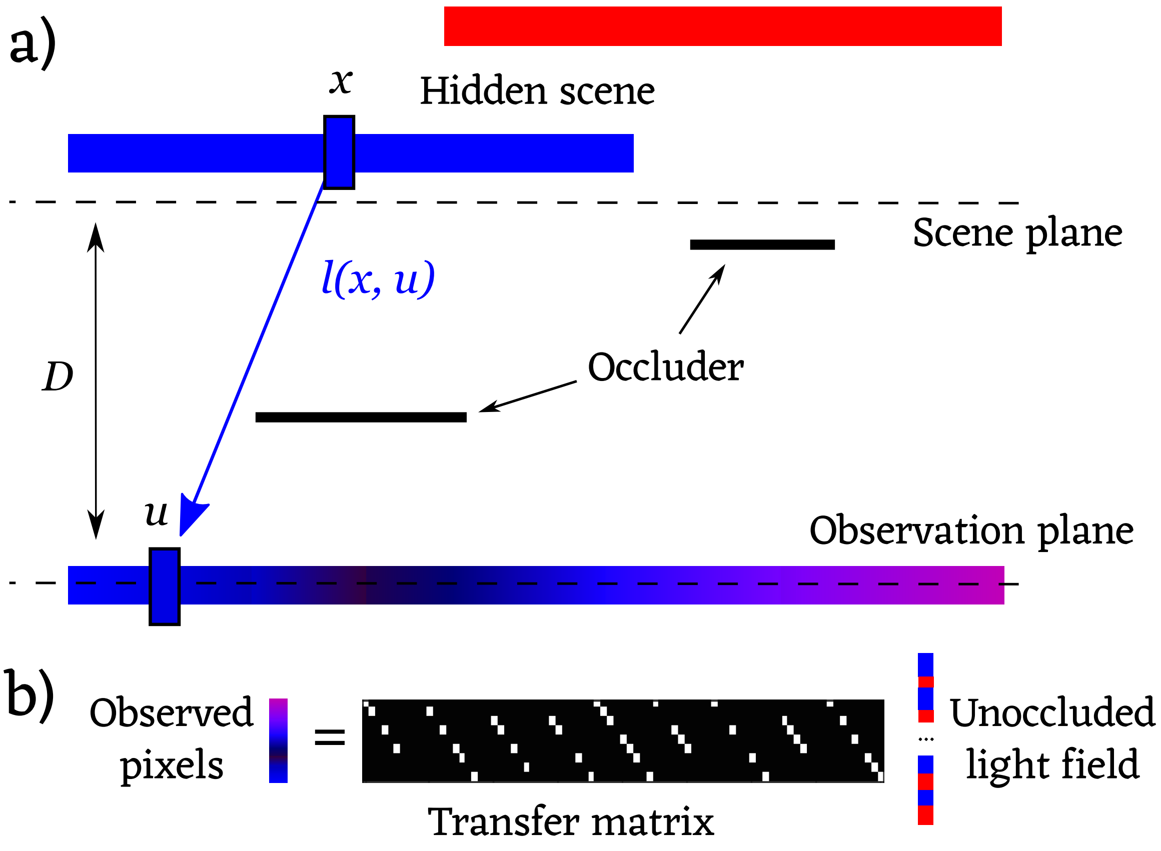


Figure 4-16: a) Simplified 2D scenario, depicting all the elements of the scene (occluder, hidden scene and observation plane) and the parametrization planes for the light field (dashed lines). (b) Discretized version of the scenario, with the light field and the observation encoded as the discrete vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The transfer matrix is a sparse, row-deficient matrix that encodes the occlusion and reflection in the system.

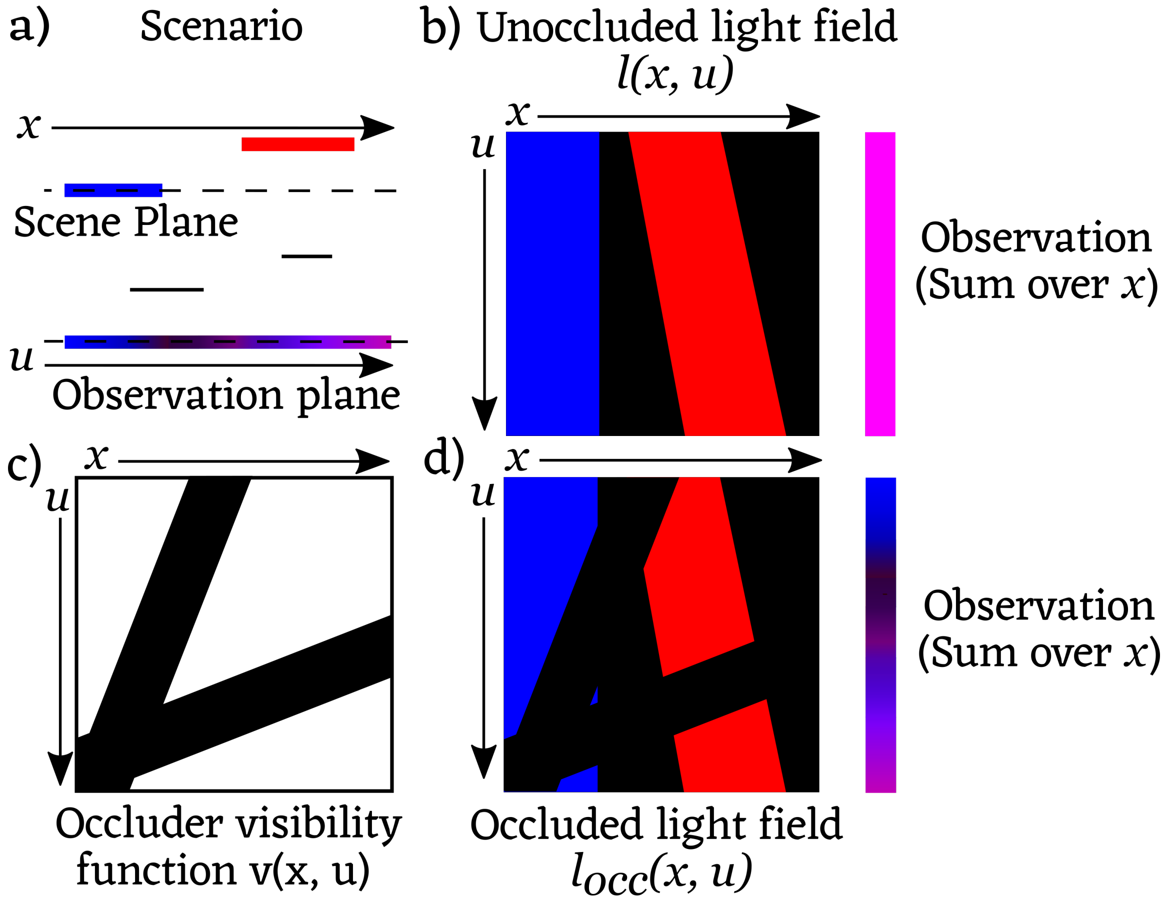


Figure 4-17: a) Sketch of a 2D imaging scenario. (b) Resulting unoccluded light field function $l(x, u)$ and observation. (c) The occluder visibility function $v(x, u)$. (d) Resulting occluded light field function $l_{occ}(x, u) = v(x, u)l(x, u)$ and observation. Note that the unoccluded observation is almost constant in u, which is not true of the occluded observation; the presence of the occluder makes the problem better-conditioned. Figure originally from [12].

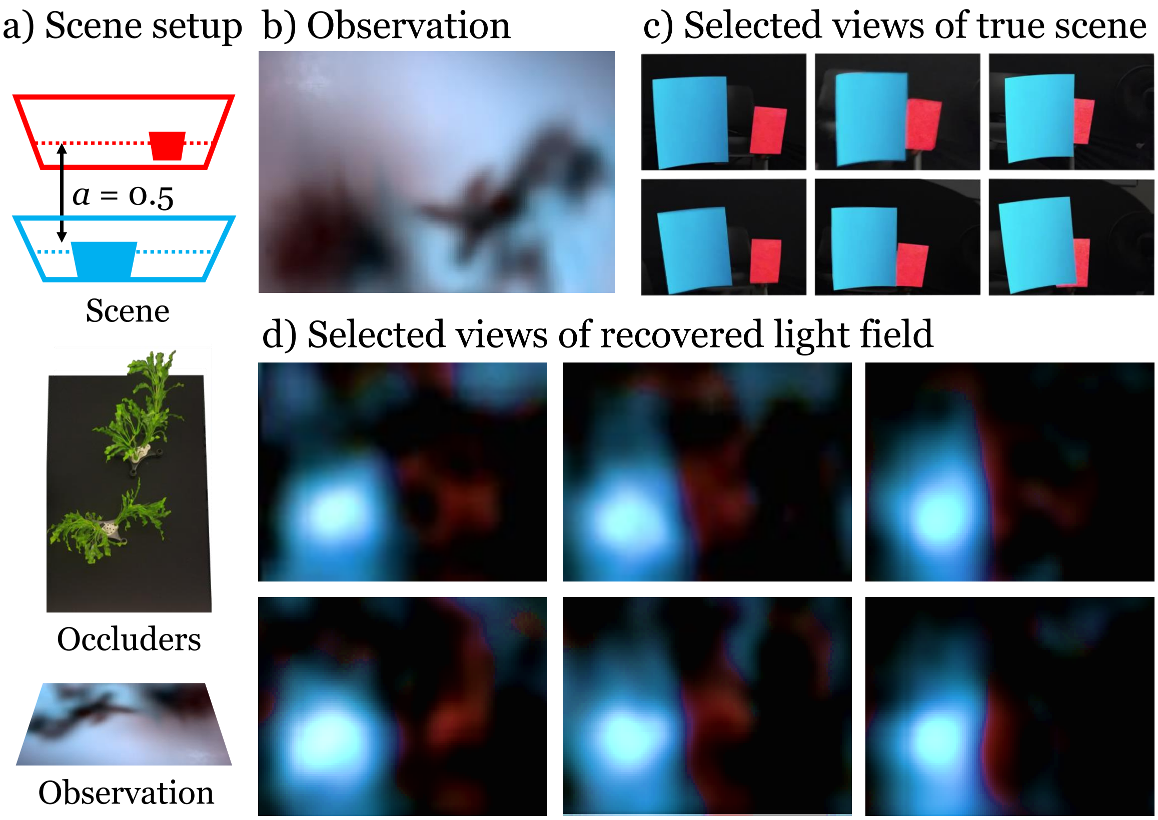

a) Scene setup
b) Observation
c) Selected views of true scene
a = 0.5
Scene
d) Selected views of recovered light field
Occluders
Observation

Figure 4-18: Reconstructions of an experimental scene with two rectangles. (a) Schematic of the setup. (b) Observation plane after background subtraction. (c) Six views of the true scene, shown in order to demonstrate what the true light field would look like. These are taken with a standard camera from equivalent positions on the observation wall. (d) Reconstructions of the light field for these views. The blue and red targets measure 8 × 12in and 6 × 8in. Figure originally from [12].

4.5 Blind Deconvolution

Most of the content in this section is taken from *Using Unknown Occluders to Recover Hidden Scenes* [110]. This section reproduces most of the paper, but includes a few additional sections, in particular, a section about alternate attempts to solve the same problem that we tried but didn't work (Section 4.5.15) and a section (Section 4.5.16) about the modifications made to the algorithm to make it run "online" (i.e. in real time).

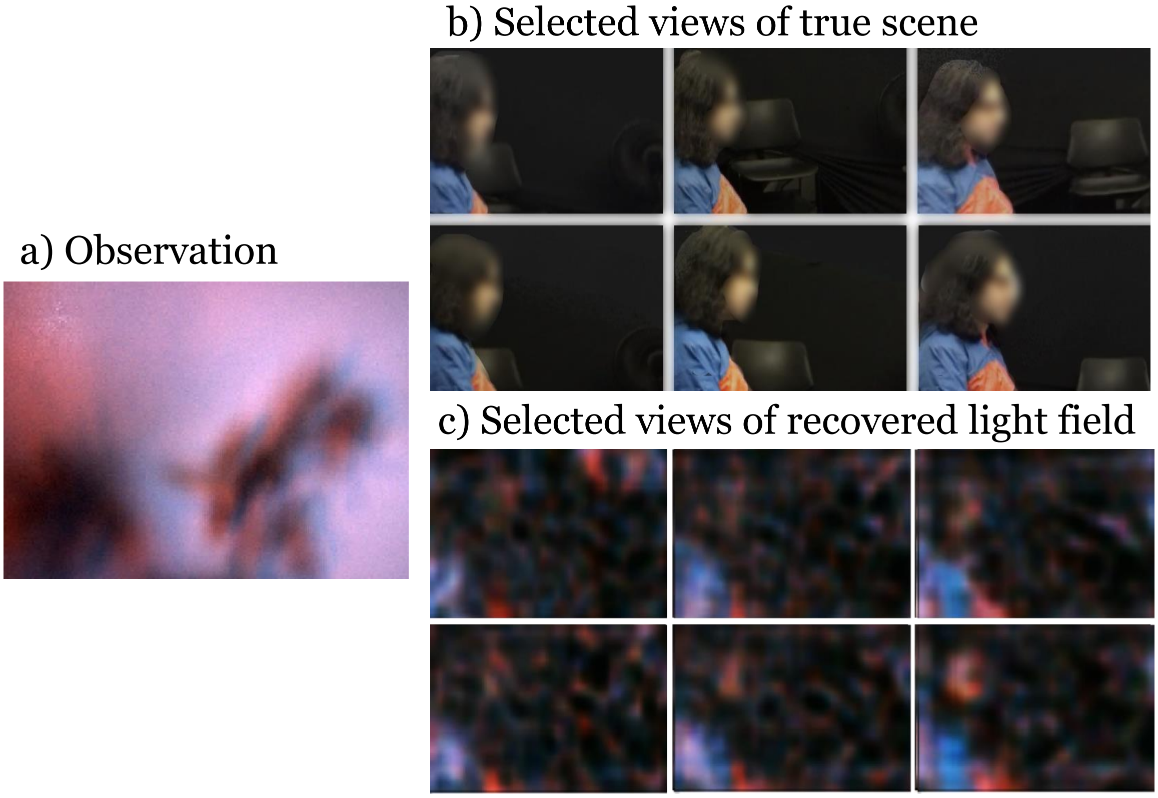

a) Observation
b) Selected views of true scene
c) Selected views of recovered light field

Figure 4-19: Reconstructions of an experimental scene with a seated subject at the scene plane. (a) Observation plane after background subtraction. (b) Six views of the true scene.(c) Reconstructions of the light field for the same views as in (b). Figure originally from [12].

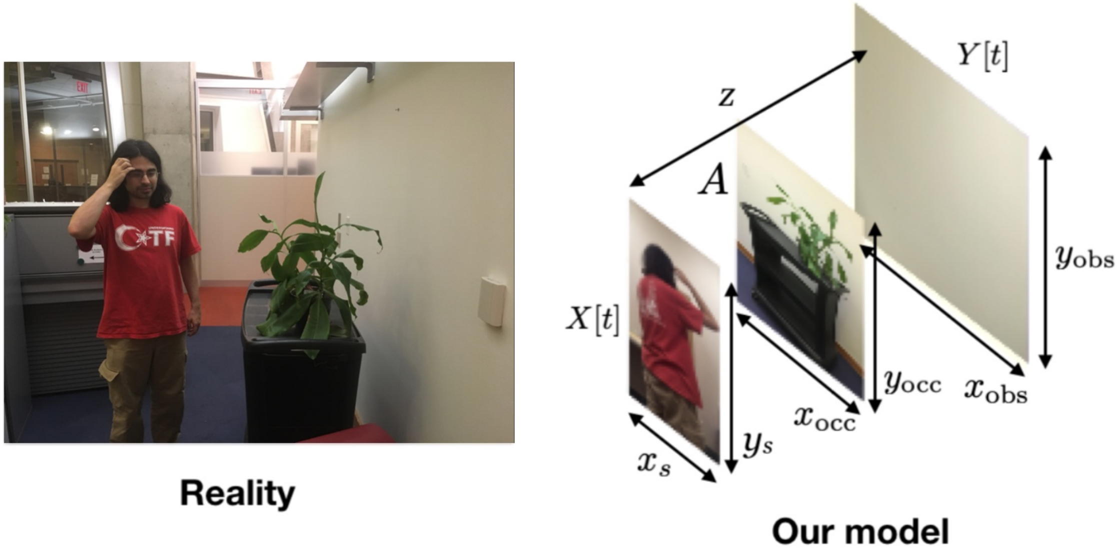


Figure 4-20: Left: a real-world scenario with a moving scene, an occluder, and an observation wall. Right: our model of the scenario.

4.5.1 Scenario

Our model of the scenario consists of three elements: a hidden moving scene, an occluder, and the observation plane. We model each of these elements as parallel 2D planes. See Fig. 4-20 for an illustration.

The hidden scene is presumed to be a collection of diffuse reflectors, shining light uniformly in all directions and towards the occluder and observation plane. The hidden scene is also presumed to contain some motion. The unknown occluder is presumed to be a set of perfectly opaque objects lying on a common plane. We assume the hidden scene, unknown occluder, and observation planes to each be a substantial distance apart, relative to their sizes. This allows us to invoke paraxial imaging assumptions, like in [109, 16].

The observation plane is presumed to be perfectly Lambertian. In simulations, we also presume the observation plane to be white and uniform, and that all of the light reaching the observation plane comes from the scene; in the experiment, we use mean-subtraction to account for non-white, non-uniform observations with ambient "nuisance" light sources, a method also employed in other work (e.g. [16]). This allows us to apply our method to most realistic scenarios with minimal adaptations to the core algorithm. We explore the effect of other deviations from the idealized scenario we present here in Section 4.5.8.

4.5.2 Light Propagation

The assumptions we describe in Subsection 4.5.1 imply that translating a light source in the scene will correspond to a simple translation of the shadow it casts on the observation plane in the opposite direction. For a more detailed explanation of why that is, and how that model deviates from reality when those assumptions are violated, see Section 4.5.8.

We model the propagation of light through the system as a 2D convolution of the scene with the occluder. This follows from the fact that a translation of an impulse light source will simply translate the shadow cast by the occluder, and from the fact that the observed light can be modeled as a linear combination of light emanating from different sources in the scene. See e.g. [109], who use the same convolution-based model of light propagation that we do. In Section 4.5.8, we go into some detail on how robust this model is, and in Section 4.5.12 we present the results of experiments, including real-world experiments.

In the simulations presented in this paper, we assume that we see the full convolution of the scene and the occluder on the wall. If the scene is a plane of size $x_s \times y_s$ and the occluder a plane of size $x_{\text{occ}} \times y_{\text{occ}}$, this corresponds to an observation of $(x_s + 2x_{\text{occ}}) \times (y_s + 2y_{\text{occ}})$. However, in practical settings, it may not be possible to see the full convolution of the scene and occluder on the wall. In the experimental case, therefore, we express the size of the observed part of the wall as $x_{\text{obs}} \times y_{\text{obs}}$. It is easy to adapt our algorithm to the case when only part of the convolution between scene and occluder is visible, as explained in Sections 4.5.3 and 4.5.7. But of course, the larger x_{obs} and y_{obs} are, the more information about the hidden scene will be available, and the better the reconstructions will be.

4.5.3 Occluder Estimation

Our blind deconvolution algorithm consists of two steps. The first step estimates the 2D occluder from the observation movie and is the primary contribution of this paper. We describe this step in this section. In Section 4.5.7, we describe the more standard second step, which recovers the movie using the estimated occluder.

4.5.4 Preliminaries

Let $Y = \{Y[0], \ldots, Y[T]\}$ be the observed video, with each $Y[t]$ corresponding to a video frame. Let $\bar{Y} = \frac{1}{T}\sum_t Y[t]$ be the "average frame" of the video, i.e., the frame such that each of its pixels is equal to the temporal average of that pixel across the entire video. Also, consider: (a) the *mean-subtracted video* $Y_\mu = \{Y[0]-\bar{Y}, \ldots, Y[T]-\bar{Y}\}$, i.e., the video of differences from the mean of the original video; (b) the *difference video* $Y_D = \{Y[1] - Y[0], \ldots, Y[T] - Y[T-1]\}$, i.e., the discrete temporal derivative of the observed video. Similarly, let $X = \{X[0], \ldots, X[T]\}$ be the ground-truth video of the scene, and let X_μ and X_D be the mean-subtracted ground-truth video of the scene and the discrete temporal derivative of the ground-truth video, respectively. X_μ and X_D are defined relative to X in the same way as above.

At this point, note the subtle but important difference between the mean-subtracted video and the difference video, which will play different roles in the algorithm. We use the observed difference video to estimate the occluder, and we use the observed mean-subtracted video when reconstructing the moving scene. What makes the difference video preferable for occluder estimation is the fact that most realistic moving scenes have just a few moving objects in them; thus each frame of the difference video is sparse.[1]

Finally, we let A be the occluder. Each element of the occluder is either 0 or 1, with 1 being no occlusion, 0 being occlusion.

[1] The sparsity of the difference video is necessary for our algorithm to work. Note, however that taking temporal derivatives amplifies the noise relative to the signal. Therefore, in situations in which the mean-subtracted ground-truth video is sparse, it is preferable to use the mean-subtracted observation video instead of the difference observation video for the task of occluder estimation. Sparse mean-subtracted ground-truth video would occur for example when most of the light in the

As explained in Section 4.5.1, we can express the observation as the convolution of the scene and the occluder[2]. Thus, for all $0 \leqslant t \leqslant T$,

$$Y[t] = A * X[t], \quad Y_\mu[t] = A * X_\mu[t], \quad Y_D[t] = A * X_D[t]. \tag{4.1}$$

Given Y (and by extension Y_μ and Y_D), our goal is to learn both X and A. We will exploit the fact that each of the $X_D[t]$ is sparse and the fact that A is binary-valued. Next, we describe an algorithm that uses $Y_D[t]$ to infer an estimate of A, which we denote $\hat{A}$.

4.5.5 Algorithm Description

Informally, we estimate the occluder by successively multiplying together randomly-chosen difference frames of the observation video with each other. Before doing so, we want to shift them such that their dot product is maximized. We can efficiently compute the set of all possible dot products of two frames, up to shifts, by computing the correlation between the two frames. Thanks to the sparsity of the difference frames, the aggregated overlap between the random difference frames that we choose will likely correspond to the shape of the occluder.

The algorithm's pseudocode is given as Algorithm 1. Therein and onwards, we use superscripts (such as $X^{i,j}$) to denote a single pixel of an image or a single entry of a matrix, and single bars (such as $|X|$) to denote the elementwise absolute value of a matrix. Note that the occluder estimate $\hat{A}$ evolves over the course of the algorithm. For the reader's convenience, we provide a detailed illustration of the first two rounds of the algorithm in Fig. 4-21.

Our algorithm consists of three steps which we repeat until a maximum iteration count is reached. First, there is a pre-processing step, the goal of which is to select

scene is being emitted by a single source.

[2] In the case that the observation is in color, Equation 4.1 will be true for each color channel individually. Then, we can run the same algorithm as otherwise, but choosing at each step a single color channel of each difference frame.

observed difference frames corresponding to sparse ground-truth difference frames. In particular, Algorithm 1 performs the preprocessing step by randomly selecting frames. We have empirically observed that this simple solution produces satisfactory results.

Next comes the alignment step. In the first iteration, we randomly select an absolute difference frame to be our first estimate of the occluder $\hat{A}_1$. At each iteration $k = 2, \ldots$ that follows, we treat $\hat{A}_{k-1}$ as a video frame which we align with a randomly selected new frame to obtain a refined estimate of the occluder $\hat{A}_k$.

In order to better understand the details of the alignment procedure and the reason why it yields an estimate of the occluder, it is instructive to consider the simple example of "ideally sparse" frames. Suppose we had a difference ground-truth frame that was a perfect impulse at (i_1, j_1), i.e., $X_D[1]^{i,j} = \delta(i - i_1, j - j_1)$, where δ is the 2D Kronecker-δ function. Then, clearly, $Y_D[1]$ is nothing but a shift of the occluder A by (i_1, j_1). In this ideal case, we immediately obtain a good picture of the occluder just by looking at a single difference observation frame. Unfortunately, in practice ground-truth video frames are only approximately sparse. We therefore model the difference observation frames as noisy shifts of the occluder. In particular, for two such frames let $Y_D[1]^{i,j} = A^{i-i_1,j-j_1} + n_1$ and $Y_D[2]^{i,j} = A^{i-i_2,j-j_2} + n_2$, where n_1 and n_2 denote noise. The goal of the alignment step is to create "aligned" versions of $Y_D[1]$ and $Y_D[2]$, which we will call $Z_D[1]$ and $Z_D[2]$, and for which:

$$\begin{aligned} Z_D[1]^{i,j} &= Y_D[1]^{i,j} = A^{i-i_1,j-j_1} + \tilde{n}_1, \\ Z_D[2]^{i,j} &= Y_D[2]^{i-(i_2-i_1),j-(j_2-j_1)} = A^{i-i_1,j-j_1} + \tilde{n}_2. \end{aligned} \tag{4.2}$$

This is achieved in Algorithm 1 by cross-correlating $Y_D[1]$ and $Y_D[2]$, finding where the max of the correlation occurs and appropriately shifting the original frames. This will approximately minimize the noise terms $\tilde{n}_1$ and $\tilde{n}_2$. See also Figure 4-21.

The goal of the third step is to reduce the noise terms in (4.2) and improve the estimate of the hidden A matrix. The simplest de-noising rule would be to return the average of $Z_D[1]$ and $Z_D[2]$. We have found instead that performing the average on the logarithms of the absolute values of the frames performs better. This explains the “geometric-mean" step in Algorithm 1.

Running the full occluder-estimation algorithm by sampling 100 frames takes a few minutes on a laptop.

Algorithm 1 Our algorithm for estimating the occluder.

```
Set $t$ to a random integer in $[0, T]$. Set $\hat{A}_1 = |Y_D[t]|$.
for $k$ in $[2, \text{NumIter}]$ do
    Set $t_k$ to a random integer in $[0, T]$.
    Compute $C = \hat{A}_{k-1} * |\overline{Y_D[t_k]}|$.
    Find $(i, j) := \text{argmax}_{i,j} C^{i,j}$, where by convention we take $C^{0,0}$ to be the central element of $C^{i,j}$.
    Let $S_{i,j}(|\overline{Y_D[t_k]}|)$ be $|Y_D[t_k]|$ shifted horizontally by $i$ pixels and vertically by $j$ pixels.
    $\hat{A}_k := (\hat{A}_{k-1})^{(k-1/k)} \cdot (|\overline{Y_D[t_k]}|)^{(1/k)}$.
    ▷ In the line above, the superscripts denote elementwise exponentiation.
    Crop the zero-valued entries of the resulting $\hat{A}_k$ until it is the same shape as $\hat{A}_{k-1}$.
end
```

4.5.6 Comparison to other methods

It is instructive to describe the differences between our application and that of most of the previous literature in blind deconvolution. Past work in blind deconvolution has largely focused on applications in image deblurring [68, 26, 59, 22]. Typically, this means that, given a single blurry image taken with a shaky camera, we would like to express the blurry image as the convolution of an unknown sharp image and an unknown blur kernel.

This problem differs from ours in three ways. First, unlike in our problem, one can assume that the unknown blur kernel is not only sparse but localized to a small region. Second, in our problem, we have additional information about the occluder: in particular, we assume it to be binary-valued. Finally, in our problem, we have many frames, each of which is a different (unknown) sparse kernel convolved with the occluder, which gives us much more information to work with.

The first difference means we have many more potential degrees of freedom to handle in our reconstruction algorithm; local search algorithms, used for deblurring

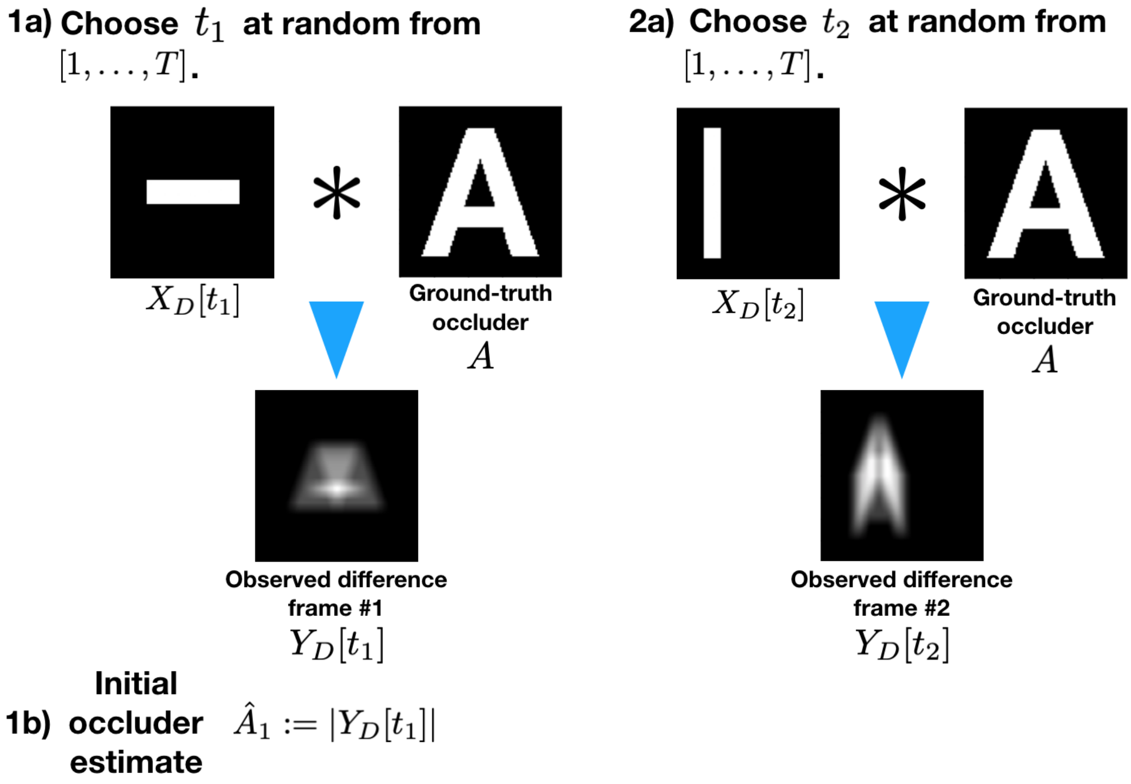

1a) Choose t_1 at random from $[1, \ldots, T]$.
$X_D[t_1]$
*
Ground-truth occluder A
Observed difference frame #1 $Y_D[t_1]$
1b) Initial occluder estimate $\hat{A}_1 := |Y_D[t_1]|$
2a) Choose t_2 at random from $[1, \ldots, T]$.
$X_D[t_2]$
*
Ground-truth occluder A
Observed difference frame #2 $Y_D[t_2]$

Figure 4-21: A worked example of initialization of Algorithm 1, followed by a single pass through the for-loop. Continued in Fig. 4-22.

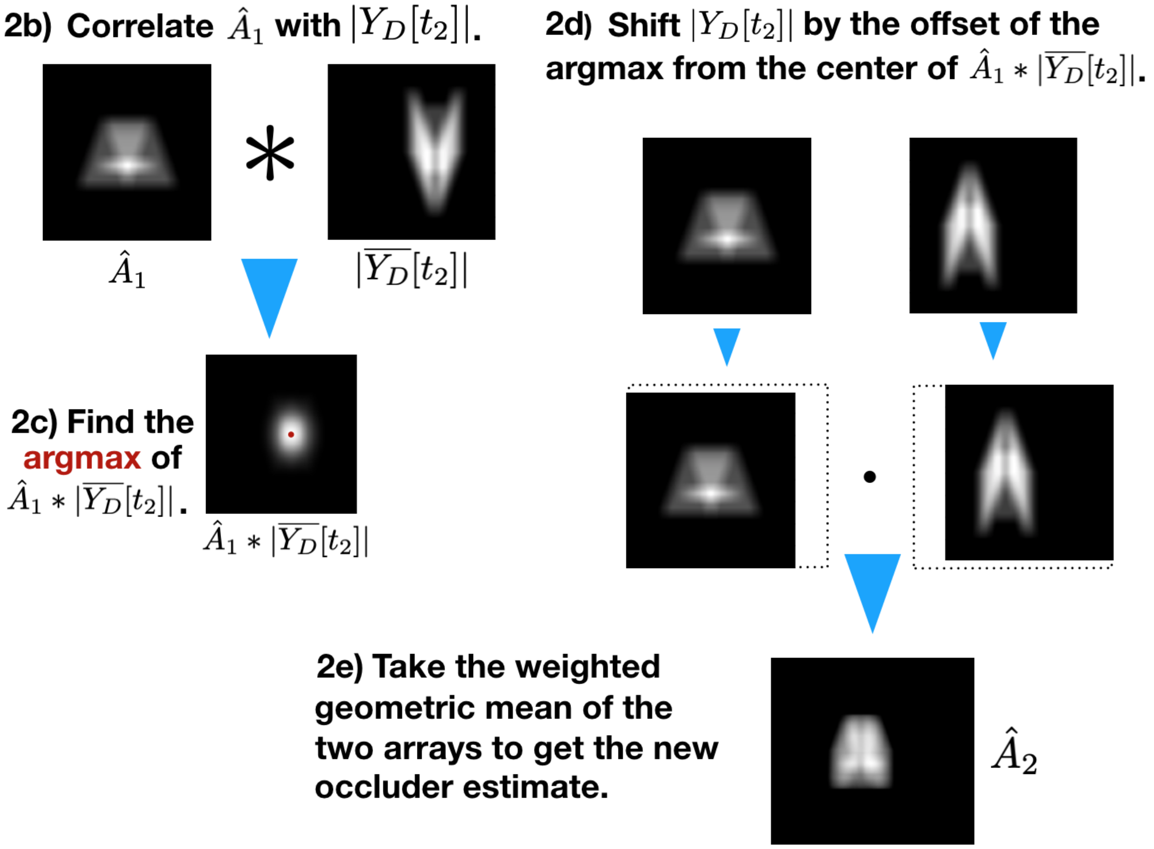

2b) Correlate $\hat{A}_1$ with $|Y_D[t_2]|$.
*
$\hat{A}_1$
$|\overline{Y_D}[t_2]|$
2c) Find the argmax of $\hat{A}_1 * |\overline{Y_D}[t_2]|$.
$\hat{A}_1 * |\overline{Y_D}[t_2]|$
2d) Shift $|Y_D[t_2]|$ by the offset of the argmax from the center of $\hat{A}_1 * |\overline{Y_D}[t_2]|$.
•
2e) Take the weighted geometric mean of the two arrays to get the new occluder estimate.
$\hat{A}_2$

Figure 4-22: A worked example of Algorithm 1.

in [26, 108], encounter difficulties when the potential size of the kernel is greatly increased. This makes it challenging to directly port blind-deconvolution algorithms used for image deblurring to our application.

Moreover, the many extra frames we have give us more information to work with. In particular, each frame shows us the occluder convolved by a different sparse kernel. This gives us many different "views" of the same occluder; it seems natural that as the length of the video goes to infinity, we should, in principle, be able to precisely characterize the shape of the occluder, even in the presence of arbitrary finite noise. How this intuition should extend to actual videos with a fixed number of frames is unclear, of course, but the nature of the problem (a fixed occluder with a non-fixed moving scene) lends itself naturally to an approach in which we estimate the occluder first, and then attempt deconvolution by the occluder estimate to recover the scene, rather than vice-versa.

Before settling on the method we used in this paper, we tried a variety of other methods, all of which failed. We tried a root-finding approach to the blind-deconvolution problem, a phase-retrieval-based approach (using ADMM), and we tried a simple gradient descent over the scene and the occluder jointly.

4.5.7 Scene Reconstruction

This section describes our method for reconstructing the moving scene, given an estimate of the occluder. In general, we reconstruct the moving scene from the mean-subtracted observation movie Y_μ[3].

To perform the reconstruction, we first formulate the matrix $\hat{\mathbf{A}}$, which describes the linear transformation corresponding to convolution by the estimated occluder $\hat{A}$. If Y_μ is of size $x_{\text{obs}} \times y_{\text{obs}}$, and the part of the scene containing movement is of size $x_s \times y_s$, then by necessity, $\hat{\mathbf{A}}$ will be a matrix of size $(x_{\text{obs}} y_{\text{obs}}) \times (x_s y_s)$[4].

[3] If the observation plane is perfectly white and uniform, and there are no "nuisance light" sources from anywhere besides the scene, the raw observation movie may be used instead.

[4] In an experimental setting, it's possible that the size of the moving scene (x_s, y_s) will be unknown. In this case, we recommend tuning the size of the scene by hand, erring on the side of larger $(\hat{x}_s, \hat{y}_s)$. If the chosen $(\hat{x}_s, \hat{y}_s)$ are too small, the reconstruction will be overconstrained and will produce nonsense; if, on the other hand, the chosen $(\hat{x}_s, \hat{y}_s)$ are too large, the expanded area will contain

Once we've formulated the forward model $\hat{\mathbf{A}}$, we can reconstruct the moving scene simply by inverting $\hat{\mathbf{A}}$ with regularization:

$$\hat{X}_\mu = \lambda(\hat{\mathbf{A}}^T \hat{\mathbf{A}} + \lambda I)^{-1} \hat{\mathbf{A}}^T Y_\mu \tag{4.3}$$

Note that in Equation 4.3, $\hat{X}_\mu$ and Y_μ have both been flattened into vectors; that is, instead of being matrices of size (x_s, y_s) and $(x_{\text{obs}}, y_{\text{obs}})$, respectively, they are vectors of size $x_s y_s$ and $x_{\text{obs}} y_{\text{obs}}$.

If we are reconstructing an RGB image, we do the calculation of Equation 4.3 for each of the three color channels individually, and then assemble them into a single image.

In Equation 4.3, the regularization parameter λ can be tuned for optimal performance. We generally found that a value of λ between 1 and 10 yielded the best reconstructions in experimental settings.

4.5.8 Deviations

In Section 4.5.1, we described assumptions that we made in order to guarantee that the observation would reflect the convolution of the occluder with the scene. For clarity, we repeat these assumptions here. First, we assume that the scene, occluder, and observation lie on parallel 2D planes. Second, we assume that the scene, occluder, and observation are far apart relative to their size. And third, we assume that the observation plane is perfectly Lambertian, white, or uniform.

Naturally, in most real-world settings, few, if any, of these assumptions will hold. So is the algorithm we present here useless in practice? No, in fact. If nothing else, in Section 4.5.12, we present the results of our algorithm in experimental settings in which all of these assumptions are violated, and these results demonstrate that our algorithm can be used in real-world settings to approximately recover hidden scenes and occluders.

We do, however, consider it instructive to describe in more detail the distortions

noise, but the subset of the scene corresponding to the signal will remain intelligible.

introduced by violating the aforementioned assumptions.

4.5.9 Non-planar or non-parallel objects

Consider the following example of an incorrect planarity assumption: suppose that we assume the occluder to be a disk, but it is in fact a sphere.

As explained in Sec. 4.5.1, the observation is the convolution of the occluder with the scene because translating an impulse light source in the scene corresponds to translating its corresponding shadow on the observation. This will be true if the occluder is a disk, but not if it is a sphere. In general, the shadow of a parallel disk on the observation plane will be a circle, but the shadow of a sphere will be an ellipse whose eccentricity will vary with the (x, y)-position of the light source. Figure 4-23 illustrates this, and shows a reconstruction of a simple scene using when incorrectly assuming the occluding sphere to be a disk.

4.5.10 Nearby objects

Assuming that the scene, occluder, and observation are far apart from each other relative to their size is common in occluder-based imaging [109, 16], and is generally called *the paraxial approximation.* The benefit of making the assumption is that it lets you ignore the effects of distance attenuation. The ray optics model tells us that if a small, flat surface of area dA is a distance r from a light source of intensity I, and the surface normal is at an angle of θ from the incident light, then the contribution c of the light source to the light intensity on the surface will go as:

$$c \sim I \frac{dA \cos(\theta)}{r^2} \tag{4.4}$$

Suppose we have two parallel planes, p_1 and p_2, a distance z apart. In that case, we can use Eq. 4.4 to derive the contribution of a light source of intensity I at $(0, 0)$ on one of the two planes to a small patch at (x, y) with area dA on the other. In this case, the contribution simplifies to:

$$c \sim I \frac{z dA}{(x^2+y^2+z^2)^{3/2}} \tag{4.5}$$

Now we can see what is meant more precisely by the scene, occluder, and observation being "far apart relative to their size." When $z \gg \sqrt{x^2+y^2}$ for all (x, y) on either plane, then Eq. 4.5 simplifies to $c \sim I\frac{dA}{z^2}$, and the contribution of a source of light at any point on p_1 to any point on p_2 will be the same, irrespective of their locations on either plane. This is a necessary condition for a translation of a light source in the scene to simply translate its observed shadow, which is in turn a necessary condition for the observation plane to reflect the convolution of the scene and occluder, as discussed in Section 4.5.1.

See Fig. 4-23 for a simulated reconstruction of a nearby scene while incorrectly assuming it to be far away.

4.5.11 Imperfections on the observation plane

Most surfaces are not perfectly uniform and white. Subtracting the mean frame from the observation video will help to reduce the effects of imperfections on the observation plane, to an extent. Color variations on the observation plane will still cause visible artifacts, however, because a darker region of the observation plane will respond less to overall increases in luminosity than a brighter region.

Non-Lambertian surfaces pose even more of a challenge. If the observation plane is sufficiently non-Lambertian, then the most important reflections off the surface will not be diffuse, but will vary strongly as a function of the angle of the incident light. This will confuse our algorithm, and probably render its output useless. However, a sufficiently non-Lambertian surface may also make the problem much easier to solve, if the observation plane is mirror-like!

We don't show a simulated example corresponding to imperfections on the observation plane in Fig. 4-23, because their effect isn't much different from simple noise, which we account for using regularization (as explained in Section 4.5.7). However, in Section 4.5.12, we show experimental results for which the observation plane includes

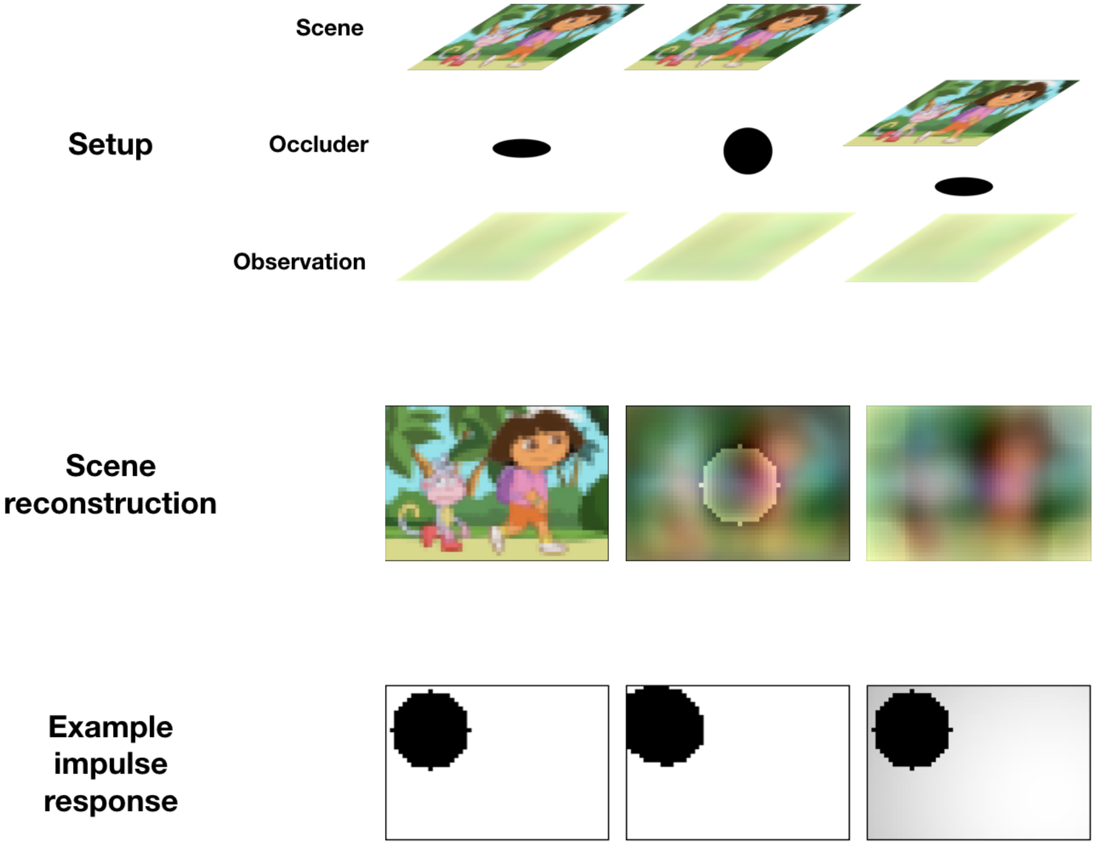


Figure 4-23: An illustration of the effects of the planarity assumption and the paraxial approximation on the reconstruction. The top row shows a sketch of the true setup; in all three cases, the assumed setup is the one on the left. The middle row shows what reconstructions, generated using the approach described in Sec 4.5.7, of the leftmost image look like when the assumptions used for that approach are violated. The bottom row shows example impulse responses for each of the three scenarios. All data shown here is simulated, with no noise, to isolate the effect of each assumption.

visible imperfections.

4.5.12 Results

In this section we present a summary of our results, both simulated and experimental. We show our reconstructions of occluders, along with a few still frames of reconstructed video. We leave the bulk of our results to the supplementary materials, however, as reconstructions of moving scenes are best seen in video form.

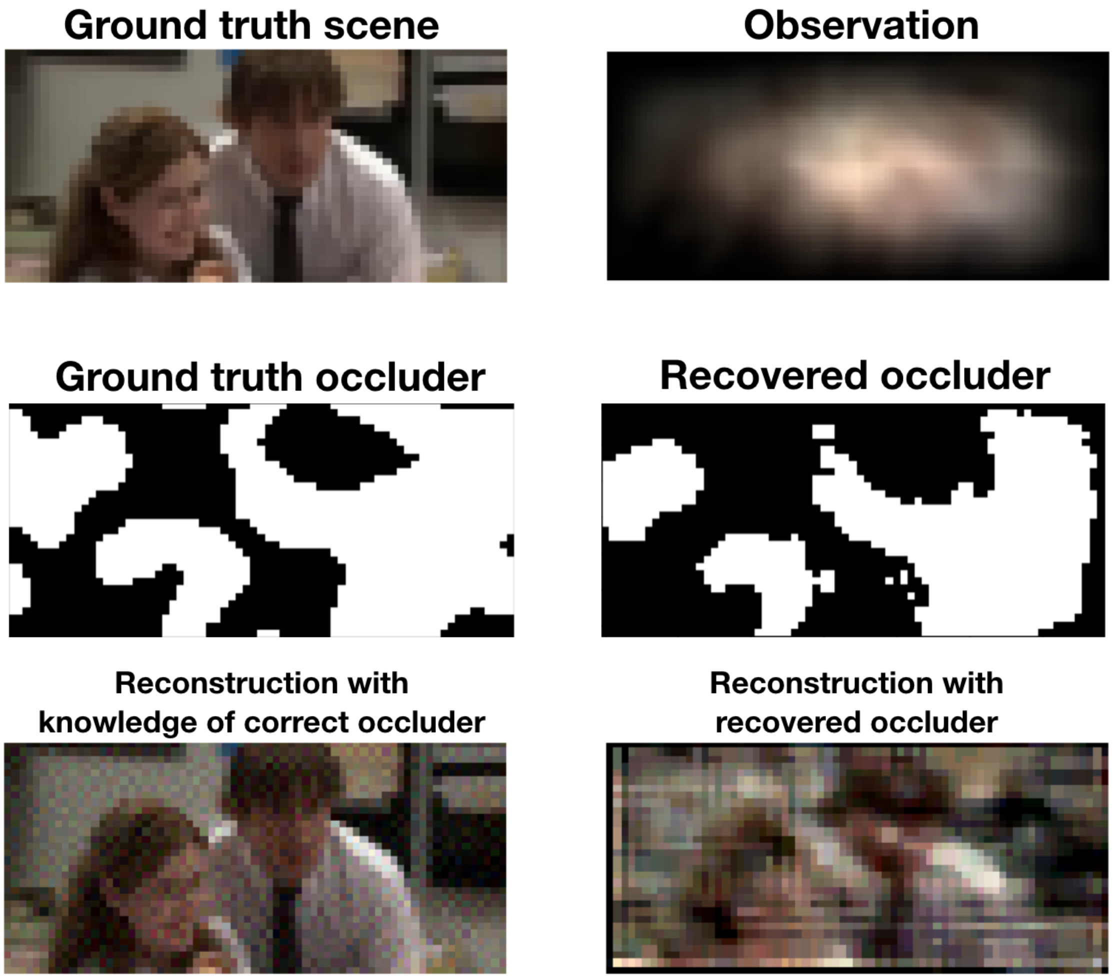

Ground truth scene
Observation
Ground truth occluder
Recovered occluder
Reconstruction with knowledge of correct occluder
Reconstruction with recovered occluder

Figure 4-24: The output of the occluder-recovery and scene-reconstruction algorithms presented in Secs. 4.5.3 and 4.5.7, using the difference frames of a simulated observation at 25dB.

4.5.13 Simulations

In this section we show the result of simulations in an ideal scenario (all of the assumptions explored in Section 4.5.8 are assumed to hold perfectly). The moving scene is the introduction to a popular television show. The ground-truth occluder was generated via a random correlated process. The observation plane is assumed to display the full convolution of the moving scene with the occluder, plus additive IID Gaussian noise. The signal-to-noise ratio on the observation plane is 25 dB.

Figure 4-24 shows the result of occluder recovery, as well as a recovered still frame from the moving scene.

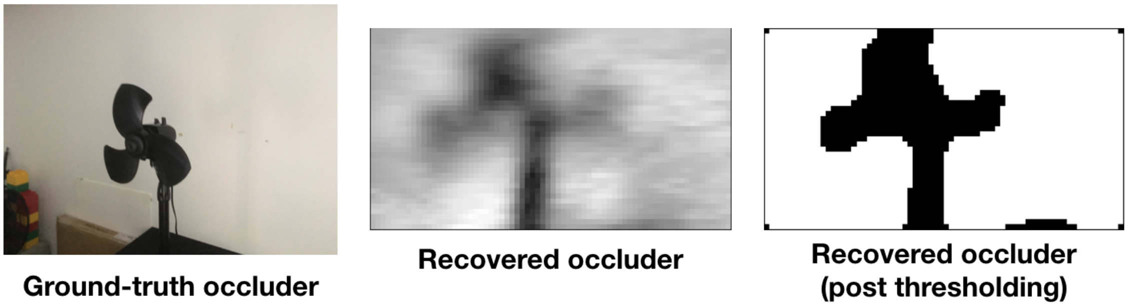
4.5.14 Experiments and Comparisons to Past Work

There has been surprisingly little past work as of this writing that does computational periscopy with the aim of recovering a head-on (as opposed to top-down) full-color 2D image of a scene in the passive setting (that is, without making use of active, directed illumination). Until 2019, the closest would have been the work of Bouman et al. in [16], but the full-color reconstructions shown in that work focus on 1D scene reconstructions, not 2D.

In 2019, however, Saunders et al. [86] showed that high-fidelity 2D full-color images could be recovered using a pinspeck occluder. Their experimental results differ from ours in two important ways. Firstly, they presume knowledge of the shape of the occluder (a pinspeck), though not its location in three-dimensional space. This gives them full knowledge of the imaging system, up to translation and scaling of the output. Second, their results are gathered from a still image with 3.5s of effective exposure time, whereas ours are drawn from a 100-FPS video of a moving scene (although the reconstruction shown is averaged over 5 frames of the ground-truth video, representing about 0.05s of exposure time). This implies a difference in signal strength. Only results from LCD monitor scenes are shown in [86]; to make it easy to compare our results with theirs, we include scenes from a cartoon shown on an LCD monitor as well as real-life scenes under heavy illumination in Figure 4-26.

As we can see in Figure 4-26, our monitor-based reconstructions are substantially lower-quality than those of [86]. We believe that this difference in reconstruction quality is primarily due to our system's imperfect knowledge of the occluder's form, which is a problem that the system of [86] does not have. We believe that the difference in SNR between the two settings may play a minor role as well.

Figure 4-25 shows the result of occluder recovery alongside its ground-truth counterpart. This recovered occluder is used for scene reconstruction in the live-action experiment in Figure 4-26.



Ground-truth occluder

Recovered occluder

Recovered occluder (post thresholding)

Figure 4-25: The output of the occluder-recovery algorithm presented in Section 4.5.3 in the experimental setting, alongside the ground-truth occluder. This is the occluder recovery used in the reconstruction shown in the second row of Figure 4-26.

4.5.15 Failed Attempts

Of course, for every good algorithm, there are any number of bad algorithms. However, the reviewers of [110], seeing that the algorithm we used was unconventional, asked in their comments why we didn't try using a few other methods instead. The answer in several cases was that we had tried them first, but found them insufficient to solve the problem. It's worth going over them, so that the reader is aware of them, both to be aware that they failed in our case but also in case they think they can make them work. Especially in the case of the phase-retrieval-approach, there may well be a possible algorithm that proves effective.

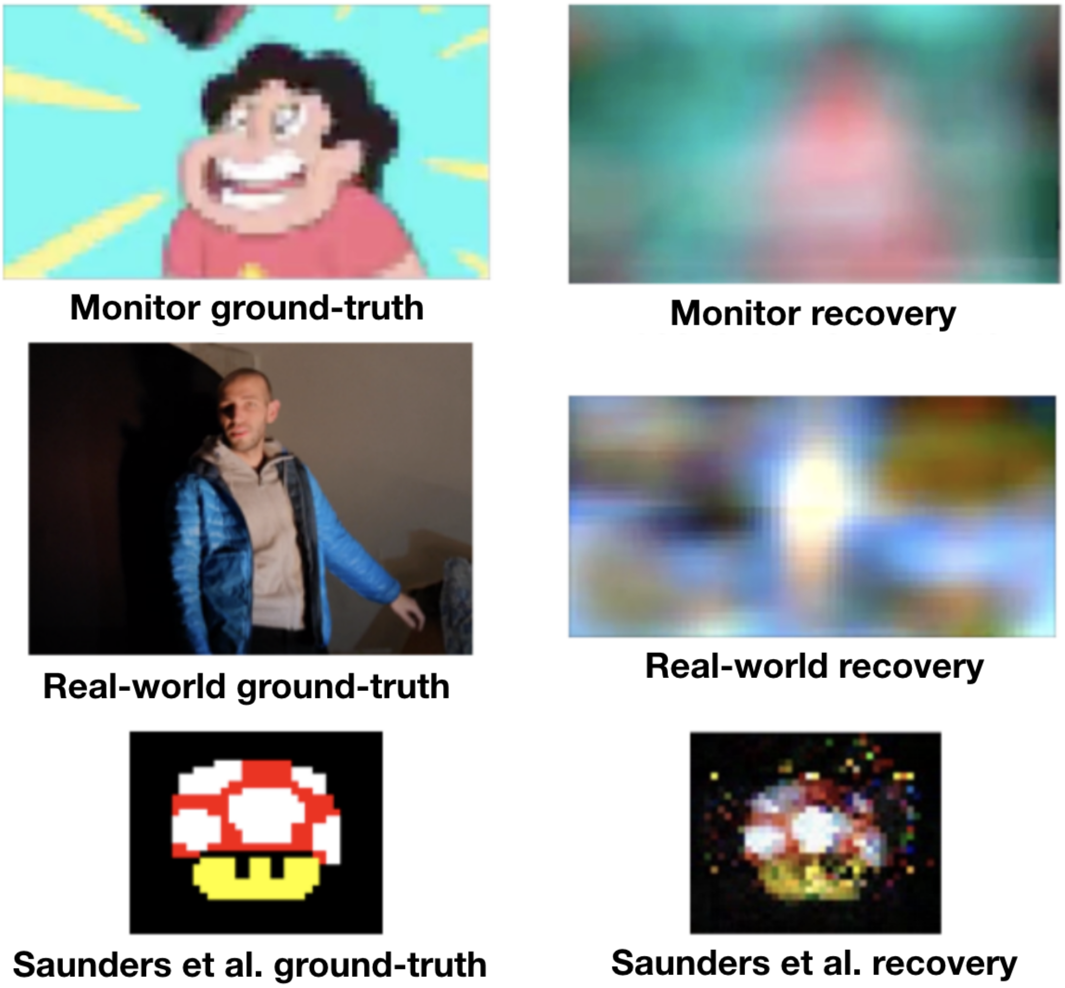
Failed Alternate Approach: Phase Retrieval

At first blush, a phase-retrieval algorithm seems like a natural approach to solving the problem at hand. After all, one can straightforwardly use the full observed movie to get a good estimate of the frequency magnitudes of the occluder, as follows.

Let Y, X, and A be the observation movie, ground-truth movie, and occluder, as before. Moreover, let $\tilde{Y}(\omega)$, $\tilde{X}(\omega)$, and $\tilde{A}(\omega)$ be the spatial 2D Fourier transforms of each of Y, X, and A. By Equation 4.1, we have, for all t and ω:

$$\tilde{Y}[t](\omega) = \tilde{A}(\omega) \cdot \tilde{X}[t](\omega) \tag{4.6}$$

Suppose we also have a prior distribution on the average frequency magnitudes of



Monitor ground-truth

Monitor recovery

Real-world ground-truth

Real-world recovery

Saunders et al. ground-truth

Saunders et al. recovery

Figure 4-26: Still frames from reconstructed videos under a variety of different experimental settings. Top row: the scene is a cartoon video, playing on an LCD monitor. Middle row: the scene is a moving man, illuminated by 200W of directed lighting. Bottom: the results of Saunders et al. [86], presented for comparison. The results of Saunders et al. demonstrate the potential improvement over our result when the form of the occluder is known. See Subsec. 4.5.14 for further discussion.

natural images $\chi(\omega)$, and the ground-truth frames of X have frequency magnitudes $|X[t](\omega)|$ drawn from this distribution. Such prior distributions exist and are well-studied; examples include [62]. Then we can also write the following:

$$\lim_{T\to\infty} \frac{1}{T} \sum t = 1^{\infty} |\tilde{X}[t](\omega)| = \chi(\omega) \tag{4.7}$$

It follows that

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} |\tilde{Y}[t](\omega)\tilde{A}(\omega)| &= \chi(\omega) \\ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{\infty} |\tilde{Y}[t](\omega)|/\chi(\omega) &= |\tilde{A}(\omega)| \end{aligned}$$

Both $|\tilde{Y}[t](\omega)|$ and $\chi(\omega)$ are known; hence, we can recover $|\tilde{A}(\omega)|$. The next step is to solve a phase-retrieval problem; given $|\tilde{A}(\omega)|$ and the knowledge that $\tilde{A}(\omega)$ is binary-valued, can we recover $\tilde{A}(\omega)$ (and therefore A itself)? We found that using alternating projections, we could solve the problem for 1D sequences of length 25 or less, or 2D sequences of size 5×5 or smaller, but for occluders much larger than that, the phase-retrieval problem quickly became intractable—especially when the frequency magnitude estimates were imperfect. In particular, the search algorithm would get stuck in local optima, or would return solutions that had frequency magnitudes very close to the correct ones but would look dramatically different from the true occluder in the primal domain. This remained true even when initializing to an occluder close in the primal domain to the correct one.

Failed Alternate Approach: Greedy Search

A greedy-search approach is another natural try at solving the problem. Even beyond simple sparsity, most natural ground-truth difference movies will be heavily constrained. For example, they are likely to have low spatial variation [16]. Hence, one can initialize the estimate of the occluder and the estimate of the ground-truth difference movie to a plausible guess, and seek to minimize a penalty function. A well-chosen penalty function of this kind would likely be formulated as a weighted sum of individual penalty terms, which could include:

1. A penalty for deviations from the observation, $\lambda_1 \sum_t ||Y[t] - \hat{A} * \hat{X}[t]||$
2. A penalty for high temporal deviations (in other words, non-sparsity in the difference movie), $\lambda_2 \sum_t ||\hat{X}_D[t]||$
3. A penalty for high spatial deviations, $\lambda_3 \sum_t \sum_i \sum_j ||X^{i,j}[t] - Xi+1, j[t]|| + ||X^{i,j}[t] - Xi, j+1[t]||$

Of course, there could be other terms in such a penalty function as well.

When we tried using gradient or stochastic gradient descent to minimize this kind of penalty function jointly over the value of the occluder and the ground-truth scene movie, we generally found that the greedy search quickly got stuck in a local optimum, even with relatively favorable initialization. We think this is because important factors in the reward function, most notably sparsity in the estimated ground-truth difference movie, only appear when the occluder estimate is almost exactly right.

4.5.16 Online Modifications

I later modified the blind-deconvolution algorithm to make it usable as an online algorithm. I wanted it to be able to handle a changing or moving occluder, to rely only on recent data (in case the occluder changed) and to be able to recover from incorrectly guessing the shape of the occluder (so that it wouldn't always return a bad result forevermore).

I aimed also in this modification to make the overall algorithm simpler and faster. As such, the algorithm stores only a minimal amount of state while running, and doesn't perform the expensive self-correlation operations described in Section 4.5.3.

Rather, the core idea of the online algorithm is as follows: for each color-frame, evaluate how close the frame is to a shadow cast by something approximating an impulse light source. (By a color-frame, I mean one color channel of a single recorded frame. So one recorded frame yields three color-frames.) Each color frame can be recovered from either the difference video or the mean-subtracted video, depending on the setup (the difference video will do better if the SNR is very high and there is localized movement in the scene, but the mean-subtracted video will do better if the scene contains only a few objects and the SNR is worse). I recommend trying both and seeing which works better in your application.

Each color-frame C is given a penalty according to a formula f that, at a high level, rewards low spatial variation, contiguity, and well-defined boundaries (on the assumption that most real-world occluders will have these properties). In addition, each color frame's penalty increases exponentially over time, parametrized by a growth rate k_2. If the new color-frame C_{new} has a lower penalty than the previously chosen color-frame C_{old}, meaning that $f(C_{\text{new}}) < f(C_{\text{old}})k_1 e^{k_2 t}$ (where C_{old} is t frames old) a new occluder estimate $g(C_{\text{new}}$ is derived from C_{new}. The details of the scoring function $f(C)$ and how an occluder estimate is derived from a color frame are described below.

The scoring function

The scoring function $f(C)$ is in fact the sum of six penalty or reward subfunctions:

$$f(C) = \lambda_1 f_{\text{var}}(C) + \lambda_2 f_{\text{diff}} + \lambda_3 f_{\text{unif}} + \lambda_4 f_{\text{edge}} + \lambda_5 f_{\text{chunk}} + \lambda_6 f_{\text{mat}}$$

The first of these, $f_{var}(C)$, is a penalty for high spatial variation. It is given as follows:

$$f_{\text{var}}(C) = \sum_{ij} \frac{d}{di} \left(\frac{d}{dj} \left(|C_{ij}|^{\frac{1}{4}} \right) \right),$$

where $\frac{d}{di}$ and $\frac{d}{dj}$ denote discrete derivatives with respect to the pixels' x and y coordinates, respectively. $\lambda_1 = 1$.

The second of these, $f_{\text{diff}}(C)$, is a reward for high overall variation (meaning that pixels are different from the mean, which implies non-trivial occluder estimates). We have:

$$f_{\text{diff}}(C) = \sum_{ij} \sqrt{|C_{ij} - \bar{C}|},$$

where $\bar{C} = \sum_{ij} C_{ij}/(xy)$, where x and y denote the dimensions of C. $\lambda_2 = -0.1$.

The third of these, $f_{\text{unif}}(C)$, is a penalty for occluder estimates that are all-on or all-off (even if that is the correct occluder, it's unlikely to yield a good reconstruction, which renders the point moot). Let $g(C)$ be the occluder estimate that would be derived from C. If we have $\bar{C} = \sum_{ij} g(C)_{ij}/(xy)$ (w here is the fraction of the candidate occluder frame $g(C)$ that is "on," meaning transmissive to light), then

$$f_{\text{unif}}(C) = \bar{C}\log(\bar{C}) - (1-\bar{C})\log(1-\bar{C}), \lambda_3 = 100.$$

The fourth of these, $f_{\text{edge}}(C)$, penalizes occluder frames in which the a lot of the occluder is on the edge of the frame. This is because that probably means the entire form of the occluder isn't in the frame at once, making this frame unsuitable for estimating the occluder. We have:

$$f_{\text{edge}}(C) = 3\sum_{ij}(1 - g(C)_{ij})e(g(C)_{ij}) + 0.04\sum_{ij} g(C_{ij})(1 - e(g(C)_{ij})),$$

where $e(p)$ is an indicator function that returns 1 when the pixel p is on the edge of the frame. $\lambda_4 = 2$.

The fifth of these, $f_{\text{chunk}}(C)$, is a penalty on the fraction of pixels that are not part of the largest contiguous chunk of the reconstructed occluder. If we let the fraction of pixels that are on but not part of the largest contiguous chunk be r, then $f_{\text{chunk}} = r$. (If this isn't clear yet, it will hopefully become clearer after you read about the occluder estimation function $g(C)$). $\lambda_5 = 100$.

The sixth and last, f_{mat}, is a reward for having the form of $g(C)$ match C. It's given by

$$f_{\text{mat}} = \frac{C \cdot g(C)}{\sum_{ij} g(C)_{ij}}.$$

$\lambda_6 = -40$. Finally, $k_1 = 0.8$ and $k_2 = 0.997$ (meaning that a new occluder should be chosen approximately every 75 frames, since the condition for choosing a new occluder is that $f(C_{\text{new}}) < f(C_{\text{old}})k_1 e^{k_2 t}$).

The occluder recovery function

Given that a frame has been judged suitable for occluder recovery, how does the algorithm use it to recover an occluder? To keep the algorithm fast and simple, the occluder estimate is simply the color-frame, thresholded at its average value. That is, $g(C)_{ij} = 1$ if $C_{ij} > \bar{C}$, and 0 otherwise, where $\bar{C} = \sum_{ij} g(C)_{ij}/(xy)$. Then, optionally, only the largest contiguous chunk of 0's is kept in $g(C)$, with the rest set to 1.

This algorithm, while more ad-hoc than the one described in Section 4.5.3, has the advantage of working nearly as well, while working in an online setting and at much lower computational cost.

4.6 Computational Mirrors: Blind Inverse Light Transport by Deep Matrix Factorization

Soon after the publication of [110], Aittala et al. [6] presented an alternate approach to the general problem of blind non-line-of-sight imaging. While this new approach doesn't use occluders to reconstruct the scene, the problem it seeks to solve is closely related to the problem solved in [110], even if the method used to solve it is very different.

In [6], the method presented is designed to reconstruct a moving, hidden scene using reflections from a "pile of clutter," whose reflectance properties are unknown. This is analogous to the problem of [110], which is to reconstruct a hidden scene using shadows cast by an unknown occluder. The former problem is a more general version of the latter problem, however. In both cases, we are given a system that can be modeled as $Y = AX + \eta$, where Y is the observation movie, X is the hidden scene movie, η is a noise term, and A is the action of some unknown object, either a pile of

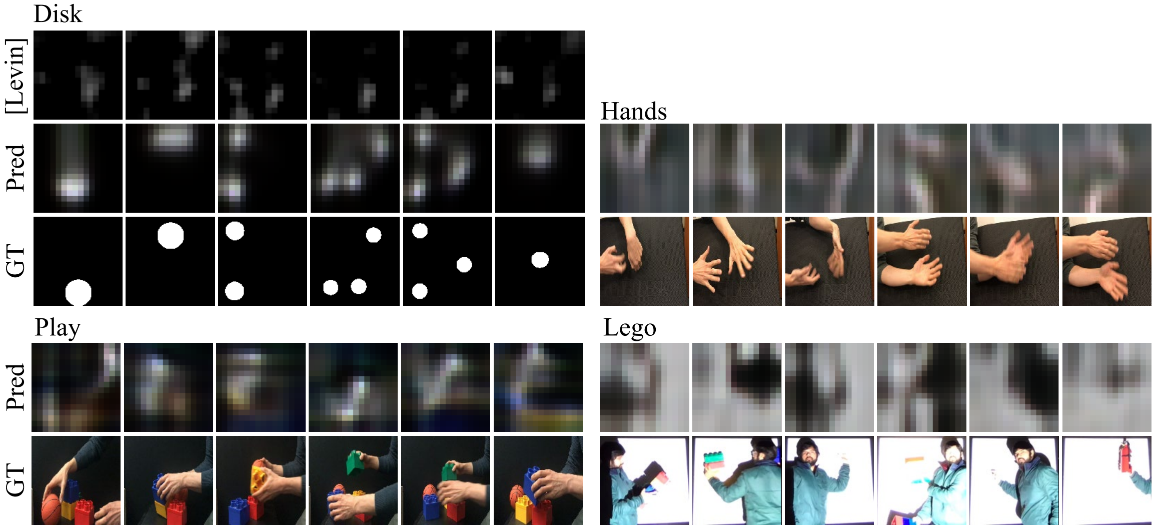


Figure 4-27: Blind light transport factorization using the method of [6]. The first three sequences are projected onto a wall behind the camera. The *Lego* sequence is performed live in front of the illuminated wall. Figure originally from [6].

clutter in the former case or an occluding object in the latter case.

Combined with the assumptions made in [110], A is known to be convolutional in the former case, but the restrictions on the form of A in the latter case are both fewer, and harder to describe. For that reason, the problem in the latter case is more difficult to solve. Accordingly, it makes use of a different method, relying on the Deep Image Prior (DIP) [100] to characterize the distribution over transfer matrices A (as well as over the scene movie). I recommend reading the paper in full if the topic is interesting, but in the interest of presenting a few of its results for comparison, I show some results from [6] in Figure 4-27.

Given the fact that the problem described in *Using Unknown Occluders* is so much narrower than that described in *Computational Mirrors*—with the requirement of an occluder-based imaging system in the first but not the second—what explains why the results are comparable in quality? After all, the later work had to contend with a completely unknown transfer matrix A, which at least in principle has many, many more degrees of freedom than a transfer matrix A that is constrained to be convolutional.

I think there are three major differences. The first is simply that the SNR was higher in the second work—the scenes were from a projector, after all. The second is that the second work's standard experimental setting is a "pile of junk" with lots of specular reflections, whereas the first not only assumes but requires a matte wall as its observation. This makes the first narrower, of course, but this time not necessarily easier; it's fairly natural that the specular reflections from shiny objects will carry a lot more information than diffuse reflections off of a matte wall. But the third reason is that I believe that the methods used in *Computational Mirrors* are better, and will likely work better in the future. Not only are they somewhat less ad-hoc, but the power of deep learning-based methods are likely only to grow as the tools and algorithms for machine learning improve. Moreover, I consider the Deep Image Prior to be a perfect fit for NLoS imaging problems in general. For that reason, I believe that those hoping to improve upon the state of the art in blind occluder-based imaging will be well-served to use the kinds of methods described in *Computational Mirrors* over those used in *Using Unknown Occluders.* And indeed, others in my academic group are working on improving upon the work in *Using Unknown Occluders* using methods similar to those described in *Computational Mirrors* now.

Additionally, see Section 5.5.1 for discussion of a third approach to this flavor of problem.

Chapter 5

Concluding Discussion

The breadth of this thesis makes it difficult to boil its conclusions down to just a few points; nevertheless, this chapter is meant to summarize the key points of the thesis, describe a tentative conclusion that could be drawn from the analysis of Chapter 3, and provide some directions for future work.

5.1 The Standard Configuration, and the Mutual Information Metric

The standard configuration as described in Section 1.4 is a useful starting point for analyses, both for optimizing occluder frames in a coded-aperture setting (see Sections 3.2 and 3.4), and for analyzing accidental-camera systems (see Sections 4.1, 4.6, and 4.5). Additionally, the mutual information metric is useful for comparing different occluders to each other. Of course, it's not the only metric we can use here (see Section 3.10 for a comparison with the MMSE metric; the two work out to do very similar things). This metric, and the surrounding analysis, is a novel contribution of this thesis.

5.2 Choosing an optimal occluder under varied conditions

Chapter 3 shed light on how to think about what the optimal occluder might be, and how to adjust for different experimental conditions. While there are a few simplified scenarios in which explicit solutions are possible (see Secs. 3.13 and 3.2.2), in most we have to rely on brute-force search or on heuristics to find the best occluder we can. Greedy searches aren't well-suited to this problem (see Section 3.3). However, it's empirically the case that choosing the best spectrally-flat mask from a menu of spectrally-flat masks at different scales performs very close to optimally, and vastly outperforms other possible methods like a greedy search; see Section 3.5.

Empirical evidence from Chapter 3 suggests that even when most of the assumptions of the Standard Configuration are relaxed, planar, spectrally-flat masks at different scales perform close to optimally. One exception is given in 3.16, where empirical evidence suggests that non-planar occluders are better than planar ones for perceiving depth.

5.3 Accidental Cameras

Chapter 4 describes accidental camera constructions under a variety of scenarios. Section 4.1 describes the simplest and most robust of the constructions, in part because it is able to use spatial averaging to increase the effective SNR of the imaging system.

Section 4.3 describes a calibrated system for reconstructing light fields. The principles at play relate closely to the discussion in Section 3.15.

Section 4.5 describes an uncalibrated system for blindly reconstructing occluders and scenes when neither is known, and Section 4.6 uses a different set of methods to attack the broader version of the problem, which includes non-occluder aperture frames. Rather than assuming a diffuse, planar observation, it actually relies on a heterogeneous observation.

5.4 Choosing a wavelength for NLoS imaging

Suppose that you want to build a non-line-of-sight occluder-based imaging system. You're not sure whether use a passive method, i.e. one that relies on light that ambient in or generated by the scene, or an active method, i.e. one that introduces its own light into the scene. You're also not sure what wavelength of light to use. You could rely on visible light, like in the NLoS systems described in Sections 4.1, 4.3, 4.5, and 4.6. You could also rely on light at other wavelengths, such as light in the infra-red (IR) spectrum, or even WiFi, as used by Adib and Katabi [3]. What are the tradeoffs inherent to each of these choices?

Imaging using IR light, even near-IR, confers a few powerful advantages. The first is that many objects of interest don't just reflect IR light, they emit it. The typical temperature of human skin is 305K. At that temperature, the peak wavelength of light emitted through black-body radiation is around 10 microns, due to Wien's displacement law [93]. A car engine will usually get up to about 330-350K, which will result in a peak wavelength of radiated light in the 8-9 micron range. In either case, this corresponds to the near-to-mid IR spectrum. If you are primarily interested in imaging objects that are generally hotter than their surroundings, like humans or cars, imaging using IR confers the great advantage that your objects of interest are generating their light at a different wavelength from their surroundings. This makes it much easier to separate the signal from the noise.

Another quasi-advantage of infra-red light is that the BRDFs of many surfaces with respect to it become closer to specular than if visible light were used. That is, a table that appears matte when illuminated with visible light may become almost mirror-like when illuminated with infra-red light. This is because matte surfaces are matte because they are jagged at a microscopic level. To oversimplify a little bit[1], it's roughly the case that if a surface looks flat at the scale of the wavelength of the incoming light, the BRDF of the surface will be approximately specular; if a surface looks jagged at the scale of the wavelength of the incoming light, the BRDF of the surface will be approximately Lambertian [52]. Thus, if a surface (as many do) looks

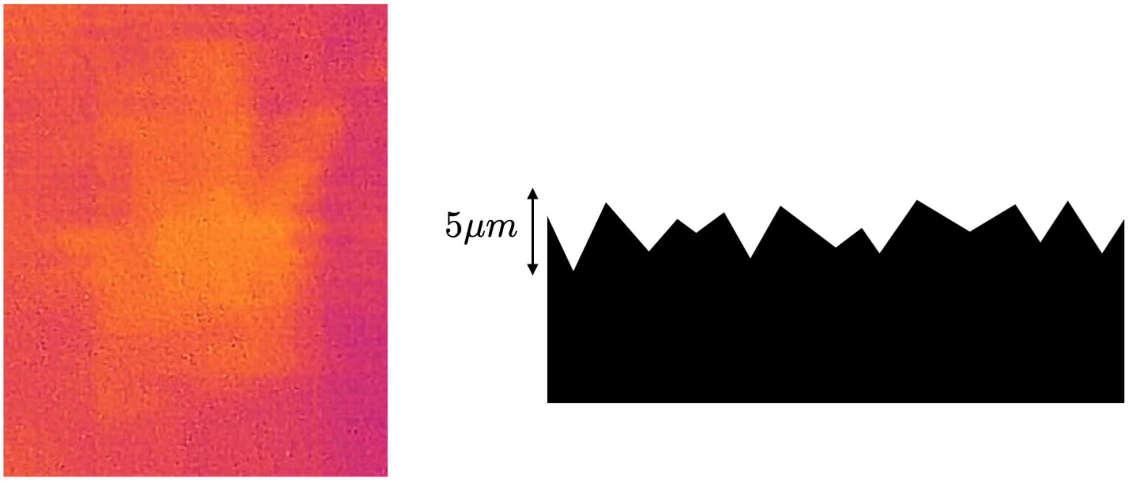


Figure 5-1: Left: an $10\mu m$ IR image of a faint reflection of a human hand, reflected by marble that appears matte in visible light. Lighter colors correspond to higher intensities. Right: a speculative depiction of the surface of the marble. Because it appears matte in visible light but is specular in IR, we can infer that it is likely jagged on a scale between $700nm$ and $10\mu m$.

matte in visible light but specular in $10\mu m$ IR light, probably it is jagged on a scale somewhere in between $700nm$ and $10\mu m$. See Figure 5-1.

I said that this is a quasi-advantage because specular surfaces often makes NLoS imaging easier, but not always. Obviously, there are plenty of situations where a specular surface makes a "NLoS imaging" problem trivial. Consider the case where you are driving a car and want to see behind you without turning your head. Looking in your rear-view mirror (a specular surface) is a great way to see objects not in your direct line of sight; no sophisticated computational techniques are necessary! However, if you are trying to see around corners described in Section 4.1, a specular floor will actually make the method unusable, while not letting you see around the corner through some other method. So while this is usually an advantage of using IR over visible light, it isn't always an unalloyed positive.

Imaging using longer wavelengths, such as $12cm$ light as used by Adib and Katabi [3], confers an additional powerful advantage: it lets you see not just around walls but

[1] This is an oversimplification because a lot of surfaces are Lambertian not only because of the scale of structures on the surface but also because of structures a little bit beneath the surface. Light with a wavelength of λ will penetrate into a surfaces to a depth of about λ, so any structures at that depth will be relevant.

through them. Light will sometimes go through opaque surfaces whose thicknesses are at the scale of its wavelength, and many walls are less than $12cm$ in thickness. This lets you solve NLoS imaging problems which would be simply impossible using shorter-wavelength light.

What is the advantage of visible light? One benefit is that most commercial cameras can record visible light, but not light at other wavelengths (though infra-red cameras have become cheaper in the last decade, with low-quality options available for a few hundred dollars or less).

Another advantage, however, is that shorter-wavelength light enables sharper reconstructions, at least in principle. When light strikes near the boundary of a surface, its probability of being reflected is proportional to the fraction of a wavelength covered by the surface. That causes the edges of a surface, seen in light of wavelength λ, to be blurred with a kernel of length approximately equal to λ, which imposes a fundamental limitation on the resolution of the image you see. The shorter the wavelength of light you use for imaging, the more this upper bound on imaging is relaxed. This resolution limit can be relevant to radio astronomy, though it usually isn't the true resolution bottleneck [33].

What about in non-line-of-sight occluder-based imaging? Is the wavelength of visible or infra-red light a bottleneck on resolution? In short, no. The notion that with current methods, you could reconstruct non-visible scenes from secondary reflections to micron or even millimeter resolution defies common sense, assuming we are trying to reconstruct macroscopic scenes. We can use the analysis of Chapter 3 to make this point more carefully. The resolution plots of Chapter 3, which give us the approximate resolution of the reconstruction we can expect if we use the best possible occluder, take as input the spatial correlation of the scene (in the form of the parameter β) and the signal-to-noise ratio.

This latter parameter is difficult to estimate directly from the brightness of various objects. It's not enough to simply divide the illumination of the observation by the scene by the illumination of the observation of nuisance objects; that will represent a dramatic underestimate of the signal-to-noise ratio, because although the signal strength is indeed the strength of the light from the scene hitting the observation, the noise comes from the variance of the light coming from nuisance objects, or camera noise. Light coming from nuisance objects (or "glare"), if it's constant and well-modeled, can simply be subtracted off. And of course, modeling error itself can be interpreted as noise.

That said, we can estimate the SNR in experimental settings empirically by analyzing the variance of reconstructed images, or difference images. Because, during reconstruction, the estimated SNR is a tunable hyperparameter, you can also get a rough estimate of the SNR by assuming it's equal to whichever value makes the reconstruction most accurate. In the experiment shown in Figure 3-41, you could reasonably estimate an SNR anywhere between 30 and 300 using these techniques. (Note, though, that this is a scene displayed on a computer monitor, with an 8-second exposure time!) Furthermore, by empirically analyzing difference frames from the observation in the problem of blind deconvolution in Section 4.5, you could estimate an SNR anywhere between 3 (from videos without additional heavy illumination from 100W lights, and with nuisance overhead lighting) to 100 (from videos with 100W lights, and without nuisance overhead lighting). Moreover, we can empirically estimate β through hyperparameter tuning; values of β between 0.3 and 0.03 yield the best reconstructions.

If we suppose that our scene and observation planes are each $2m$ to a side, and we are attempting to get a 2D reconstruction using an ideal (spectrally flat) occluder, we can adapt the method used to generate Figure 3-7 to create a new figure, Figure 5-2, which estimates the best possible resolution as a function of the SNR and β. We can see from Figure 5-2 that even under the very most optimistic conditions, using the best possible occluder we can only expect centimeter-level resolution in our reconstruction. This leads us to the conclusion that we have nothing whatsoever to fear from using IR light rather than visible light, and only under the very best of plausible conditions can we expect to be bottlenecked on the wavelength of WiFi as used by Adib and Katabi ($12cm$). All of this assumes a 2D reconstruction and a solely occluder-based approach, of course; it's possible that dramatically better

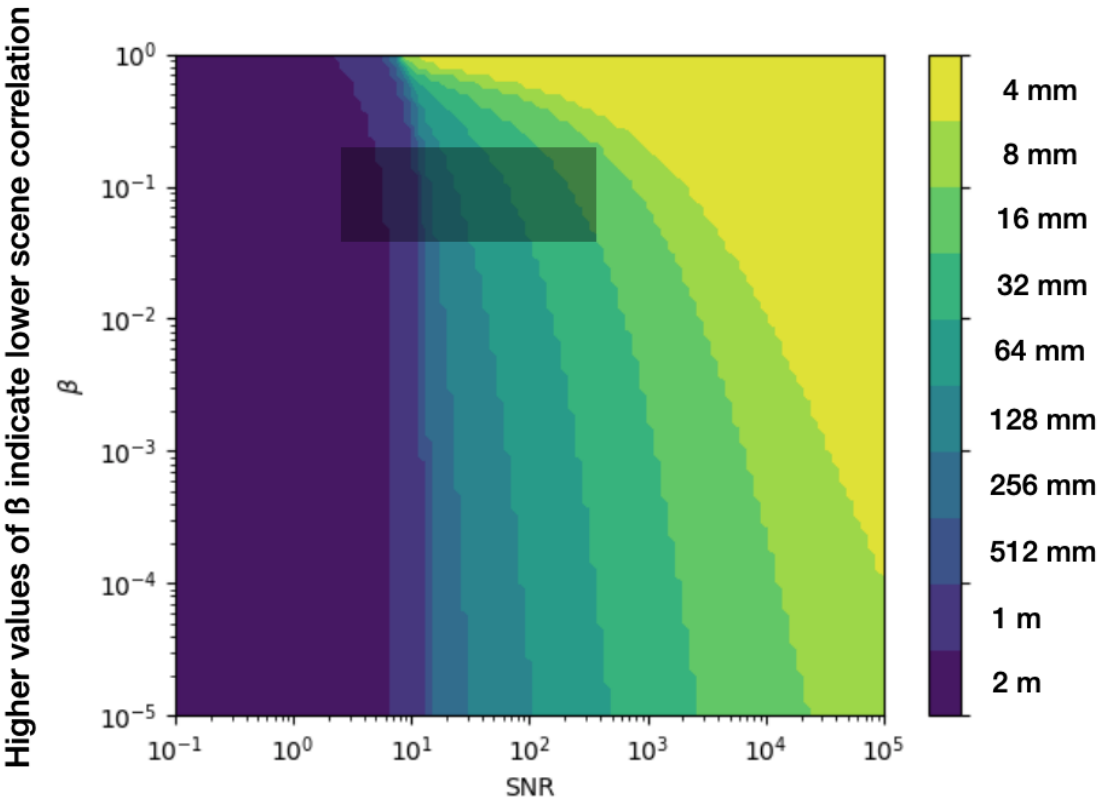


Figure 5-2: A reprint of Figure 3-7, assuming a $2m$ scene and observation. The dark rectangle indicates a range of realistic experimental settings for a 2D passive occluder-based reconstruction. As shown above, even with the ideal occluder, resolution on a scale of centimeters is very difficult to achieve.

performance could be achieved with some other approach. Until then, however, we would do well not to fear using millimeter- or even centimeter-wave light for non-line-of-sight imaging.

5.5 Future Work

This section provides a few interesting directions for future work.

5.5.1 Blind inversion as video unscrambling

In the vein of Section 4.6, there is possible approach that may prove useful, perhaps in conjunction with other methods. Its core configuration is nearly as narrow as that of [6], but it makes nearly opposite assumptions. Rather than assuming an occluder in between the scene and observation, and a perfectly matte observation plane, it assumes no occluder and a perfectly specular (but warped) observation plane.

In the language of $Y = AX + \eta$, this time our assumption is that A is an unknown permutation of the identity matrix. This sounds unrealistic, but in fact it's not necessarily far from the truth. If we pay attention only to the shiny objects around us, and they're pointing to something random, then the transfer matrix that results (after we've thrown away all the non-specular observations) will look a lot like a permutation matrix.

The problem that results is very interesting. In fact, it's something like a jigsaw puzzle. The observation movie we get is exactly the scene movie, but the pixels' order has been scrambled. How do we tell which pixel should go where? Well, the answer is that if two pixels are similar over the course of the movie, they should be near each other.

Indeed, the input to this miniature version of the blind inversion problem is just a scrambled movie, not scrambled in time but in space, so that each pixel plays the movie in sequence but from an unknown location. If we are willing to accept that pixels that have similar "histories" (i.e. look similar over the course of the movie) should be nearby, then the problem effectively becomes a dimensionality-reduction problem. Each pixel's history is a t-dimensional vector, where t is the number of frames in the scrambled movie[2]; its location in space is a 2-dimensional vector. We would like to find a 2-dimensional embedding for the movie, so that for each pair of pixels that's nearby in the high-dimensional embedding (i.e. those two pixels have similar histories), that pair of pixels is nearby in the 2-dimensional embedding (i.e. they're close to each other in the recovered movie).

There are a variety of different well-known dimensionality-reduction algorithms already implented efficiently, and any of these can be applied to the problem. While it's not a perfect fit, because they are often used to identify clusters of data rather

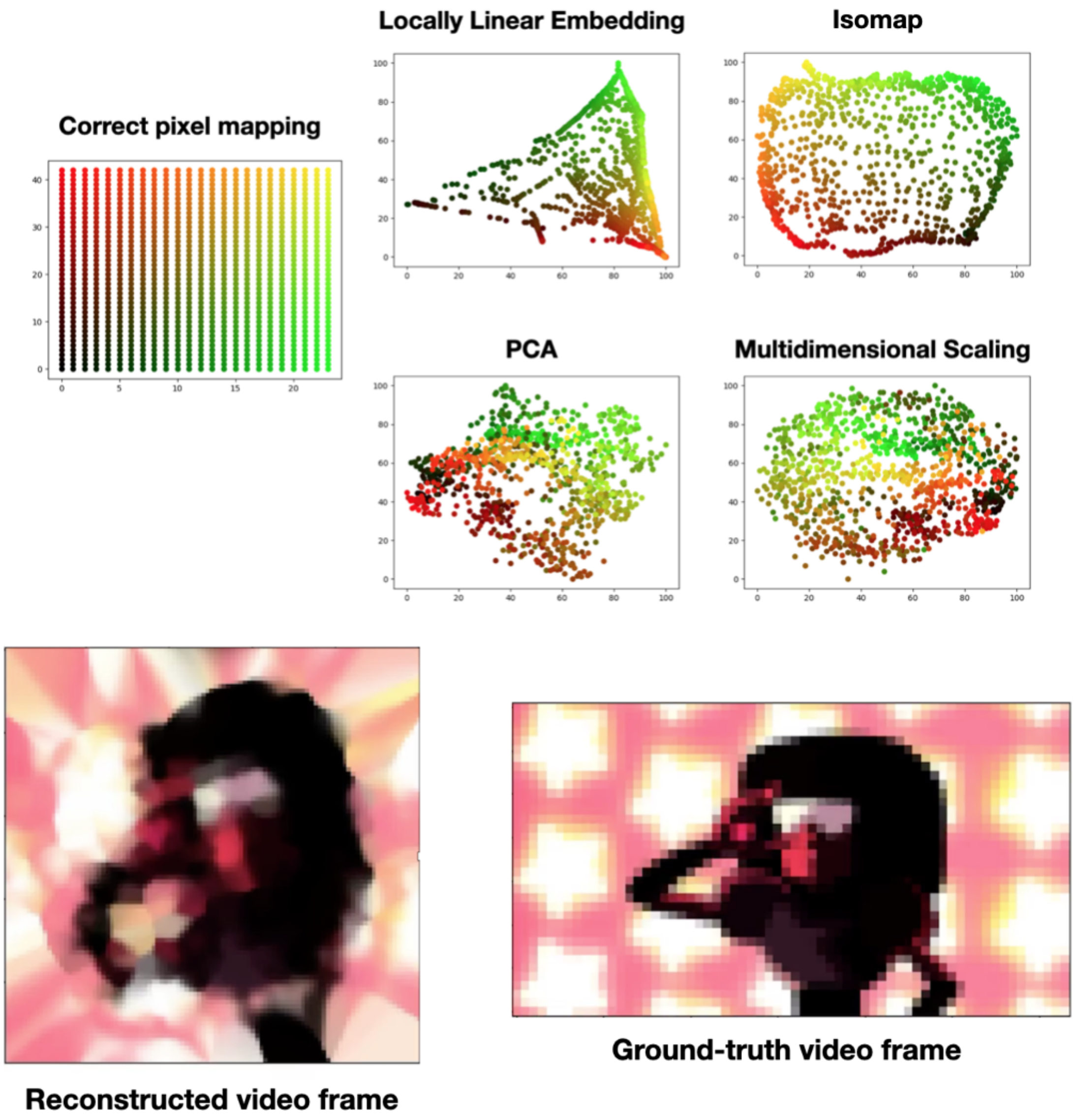
[2] In fact, in an ordinary color movie, it's a $3t$-dimensional vector; each color channel produces one number, for each frame. But this is only a minor complication.

than putting together a puzzle, most of them work well. I have had the most success using the so-called Isomap algorithm, developed by Tenenbaum et al. in 2000 [95]. See Figure 5-3 for a comparison of different stock dimensionality-reduction algorithms in simulation.

We can apply this technique to an experimental "unscrambling" problem: reconstructing a video seen only through an object that produces unknown distortions. In Figure 5-4, we view a medium-length (approx. 3000 frames) video through a water bottle, and attempt to reconstruct the true video. Before we directly apply a dimensionality-reduction method, we want to use a method to exclude pixels that don't provide much additional information, either because they're constant in value (and thus not actually transmitting light from the video) or are too similar to another pixel (and won't provide information above and beyond what that pixel is providing). This latter point is important, because of the way the bottle (and many other reflective surfaces) warps; along some dimensions, the bottle's intensity changes rapidly, whereas along others, it stays almost constant. It's a huge computational win to be able to exclude redundant pixels.

To do this, the ball-tree data structure is a very natural choice. The ball-tree allows querying the existence of other elements within a certain radius (or "ball") of a given point. In this case, what we can do is enter a new pixel color history into the ball-tree only if no other existing pixel color history has been entered into the ball-tree within a given radius (that is, no other pixel has a history too similar to the new one). In the experiment shown in Figure 5-4, it allows us to cut down from almost 500,000 raw pixels to only about 10,000 pixels.

While these results show promise for real-world applications—the hope would be to be able to reconstruct an entire scene surrounding a single irregular but specular object simply by filming the object from different vantage points—the problem becomes much harder when the video doesn't include "scene changes," as the videos in Figs. 5-3 and 5-4 did. This is because scene changes give us many different views of the same pixels, letting us learn which pixels are truly nearby; with only one scene, pixels that are part of the background and far apart, but with similar colors, will be



Correct pixel mapping

Locally Linear Embedding

Isomap

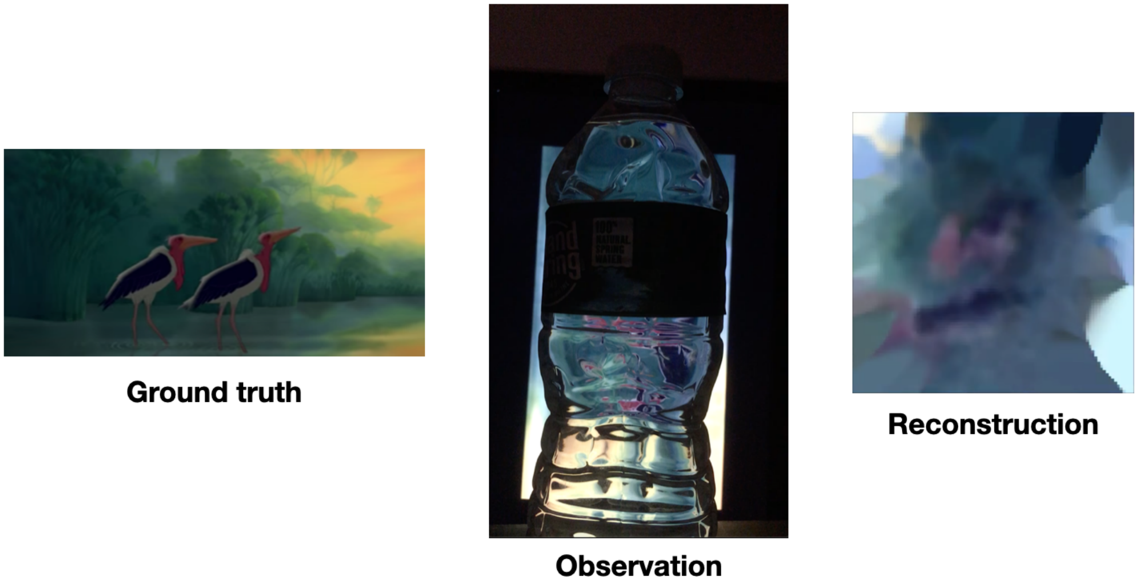
PCA

Multidimensional Scaling

Reconstructed video frame

Ground-truth video frame

Figure 5-3: Top: The unscrambling results using each of four different dimensionality reduction algorithms for a short (624-frame) video, scrambled uniformly at random, with the best results achieved by Isomap, followed by Locally Linear Embedding. Bottom: A comparison of a reconstructed (unscrambled) frame with a ground-truth frame from the original video, using a Voronoi mapping of colors to locations in the unscrambled video. As can be seen, the reconstruction is quite good, except at the edges of the frame where the scale is wrong.



Ground truth

Observation

Reconstruction

Figure 5-4: We can use dimensionality-reduction techniques and the ball-tree data structure to reconstruct a 3000-frame video seen through a water bottle. Shown is a single frame of the reconstructed video. Note that because of fundamental ambiguities, the reconstructed frame is a rotation and reflection of the ground-truth frame.

erroneously identified as near each other. With enough movement in the scene, the problem is still solvable, but becomes much more challenging.

5.5.2 Using perpendicular occlusion to gain parallax

This is an idea first put into practice in the latter half of [16], in which two edge cameras are combined at a doorway to create a parallax effect which enables the perception of depth. However, the discussion in Section 3.20 suggests that this kind of technique (of using multiple 1D occluders in tandem to infer depth) is possible whenever the occluder is perpendicular to the observation plane. Moreover, as we saw in Section 4.1, edge cameras are actually one of the least effective 1D occluding patterns that nevertheless enable full reconstruction. Might this technique be possible with other relatively ubiquitous 1D occluders?

One example of a fairly common pattern is a “signpost” camera, which could use two signposts as 1D pinspeck cameras, and use parallax to infer depth, as before, perhaps to infer the locations of cars on a nearby street. Figure 4-2 suggests that an equivalent SNR would be able to produce results at least as good as those in the latter half of [16], and one might hope cars, because of how reflective they are, could yield a better experimental SNR than the doorway camera. (Of course, at night, with headlights on, the problem becomes particularly easy!)

See Figure 5-5 for an image of what an example observation plane might look like.

5.5.3 Applying machine learning for NLoS imaging techniques

Of all the directions for future work presented here, this is certainly the most promising one, or at least the broadest. Recently, there have been several results that make use of machine learning techniques for the purpose of NLoS imaging [99, 61, 6]. Additionally, there is active research on improving upon the results of [110], which I expect will soon succeed.

Using machine learning for general NLoS imaging problems is a very natural choice, especially as the state of general-purpose machine vision algorithms contin-

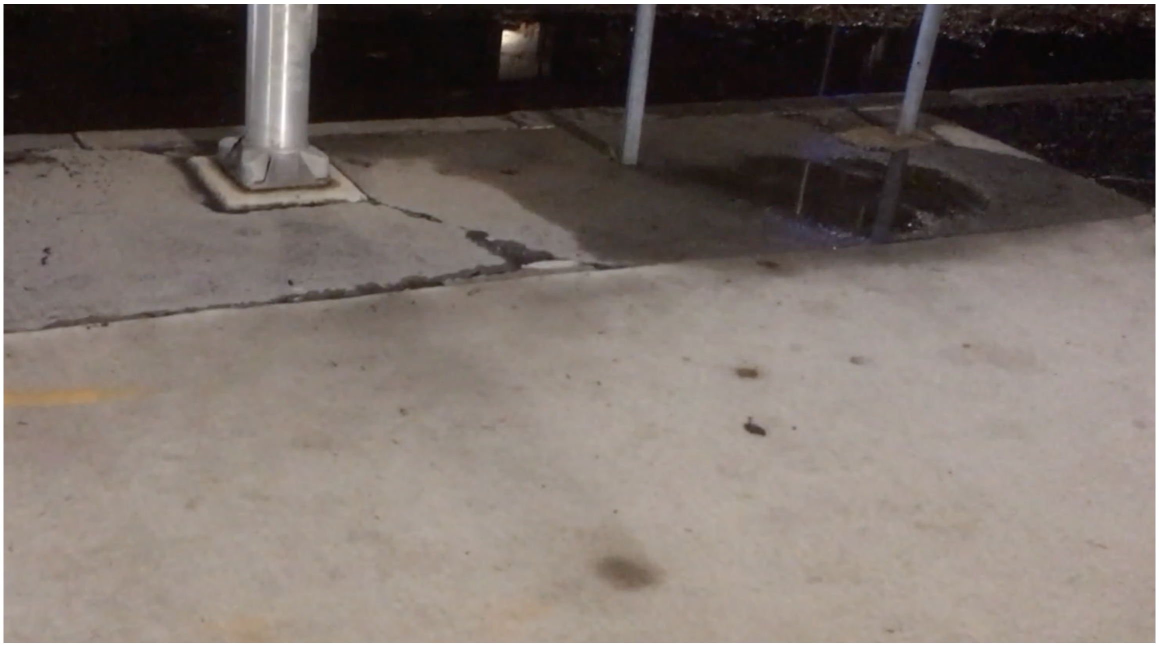


Figure 5-5: A possible observation plane that could make use of parallax from multiple 1D occluders. Each of the signpoles near the road act as 1D pinspeck occluders, but light sources in the street will cause them to cast shadows that extend in different directions, creating a potentially useful parallax effect.

ues to improve, as it has throughout this decade [60, 83, 39]. Indeed, at this point, image recognition using machine learning has arguably surpassed human-level performance [53]. Moreover, generative models for image models have drastically improved in the last few years, thanks to generative adversarial networks [81, 11, 36]. These techniques work well true even for generating highly technical images that exist only in a particular domain, such as medical images [51]. Such techniques seem like a very natural fit for generating reconstructions in the context of NLoS imaging.

More generally, while the NLoS imaging techniques described in Chapter 4 work well for their particular applications, as long as the algorithms underlying them are designed by humans and not learned by machines, they will never be able to make use of all of the information existent in the observation plane, especially in the case of uncalibrated systems like [110]. Optimistically, a machine learning system might be able to make use of information that a human designer would never think to bake into their algorithm.

The field of passive NLoS imaging, which barely existed a decade ago, will, I anticipate, flourish in the decade to come, perhaps thanks to the impressive power and generality of deep learning when applied to image problems. I look forward to seeing the results that will come next.

Bibliography

[1] JG Ables. Fourier transform photography: a new method for X-ray astronomy. *Publications of the Astronomical Society of Australia*, 1(4):172–173, 1968.

[2] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992.

[3] Fadel Adib and Dina Katabi. See through walls with WiFi! In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 75–86, 2013.

[4] Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C Sankaranarayanan. Convolutional approximations to the general non-line-of-sight imaging operator. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7889–7899, 2019.

[5] Timo Aila and Ville Miettinen. dpvs: An occlusion culling system for massive dynamic environments. *IEEE Computer graphics and Applications*, 24(2):86–97, 2004.

[6] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, and Fredo Durand. Computational mirrors: Blind inverse light transport by deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 14311–14321, 2019.

[7] Ganesh Ajjanagadde, Christos Thrampoulidis, Adam Yedidia, and Gregory Wornell. Near-optimal coded apertures for imaging via Nazarov's theorem. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7690–7694. IEEE, 2019.

[8] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. DiffuserCam: lensless single-exposure 3D imaging. *Optica*, 5(1):1–9, 2018.

[9] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard Baraniuk. Flatcam: Thin, bare-sensor cameras using coded aperture and computation. *arXiv preprint arXiv:1509.00116*, 2015.

[10] Mehdi Askari, Seong-Bok Kim, Kwang-Soo Shin, Seok-Bum Ko, Sang-Hoo Kim, Dae-Youl Park, Yeon-Gyeong Ju, and Jae-Hyeung Park. Occlusion handling using angular spectrum convolution in fully analytical mesh based computer generated hologram. *Optics Express*, 25(21):25867–25878, 2017.

[11] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

[12] Manel Baradad, Vickie Ye, Adam B Yedidia, Frédo Durand, William T Freeman, Gregory W Wornell, and Antonio Torralba. Inferring light fields from shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6267–6275, 2018.

[13] Jiří Bittner, Michael Wimmer, Harald Piringer, and Werner Purgathofer. Coherent hierarchical culling: Hardware occlusion queries made useful. In *Computer Graphics Forum*, volume 23, pages 615–624. Wiley Online Library, 2004.

[14] Arup Bose, Rajat Subhra Hazra, Koushik Saha, et al. Limiting spectral distribution of circulant type matrices with dependent inputs. *Electron. J. Probab*, 14(86):2463–2491, 2009.

[15] Arup Bose and Joydip Mitra. Limiting spectral distribution of a special circulant. *Statistics & probability letters*, 60(1):111–120, 2002.

[16] Katherine L Bouman, Vickie Ye, Adam B Yedidia, Frédo Durand, Gregory W Wornell, Antonio Torralba, and William T Freeman. Turning corners into cameras: Principles and methods. In *International Conference on Computer Vision*, volume 1, page 8, 2017.

[17] Sara Bradburn, Wade Thomas Cathey, and Edward R Dowski. Realizations of focus invariance in optical–digital systems with wave-front coding. *Applied optics*, 36(35):9157–9166, 1997.

[18] David J Brady, Nikos P Pitsianis, and Xiaobai Sun. Reference structure tomography. *JOSA A*, 21(7):1140–1147, 2004.

[19] David E Breen, Ross T Whitaker, Eric Rose, and Mihran Tuceryan. Interactive occlusion and automatic object placement for augmented reality. In *Computer Graphics Forum*, volume 15, pages 11–22. Wiley Online Library, 1996.

[20] Richard P Brent and Adam B Yedidia. Computation of maximal determinants of binary circulant matrices. *arXiv preprint arXiv:1801.00399*, 2018.

[21] Peter J Cameron. Encyclopaedia of design theory, chapter on Hadamard Matrices. *Cayley graphs and coset diagrams*, pages 1–9, 2013.

[22] Michael Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):58–63, 1976.

[23] TM Cannon and EE Fenimore. Tomographical imaging using uniformly redundant arrays. *Applied Optics*, 18(7):1052–1057, 1979.

[24] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 307–318, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[25] Edwin B Champagne. Nonparaxial imaging, magnification, and aberration properties in holography. *JOSA*, 57(1):51–55, 1967.

[26] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998.

[27] Roy C Chaney and Olga Vassilieva. Method and system for reducing background artifacts from uniformly redundant array collimators in single photon emission computed tomography, June 17 2003. US Patent 6,580,939.

[28] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A Bayesian perspective on the Deep Image Prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5443–5451, 2019.

[29] S Chowla. A property of biquadratic residues. *Proc. Nat. Acad. Sci. India Sect. A*, 14:45–46, 1944.

[30] Michał J Cieślak, Kelum AA Gamage, and Robert Glover. Coded-aperture imaging systems: Past, present and future development–a review. *Radiation Measurements*, 92:59–71, 2016.

[31] Adam Lloyd Cohen. Anti-pinhole imaging. *Journal of Modern Optics*, 29(1):63–67, 1982.

[32] Adam Lloyd Cohen. Anti-pinhole imaging. *Optica Acta: International Journal of Optics*, 29(1):63–67, 1982.

[33] Marshall Harris Cohen and Kenneth I Kellermann. Quasars and active galactic nuclei: high resolution radio imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 92(25):11339, 1995.

[34] Philip S Considine. Effects of coherence on imaging systems. *JOSA*, 56(8):1001–1009, 1966.

[35] Satyan Coorg and Seth Teller. Real-time occlusion culling for models with large occluders. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 83–ff, 1997.

[36] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[37] Edward R Dowski and W Thomas Cathey. Single-lens single-image incoherent passive-ranging systems. *Applied Optics*, 33(29):6762–6773, 1994.

[38] Marco F Duarte, Mark A Davenport, Dharmpal Takbar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

[39] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

[40] EE Fenimore. Coded aperture imaging: predicted performance of uniformly redundant arrays. *Applied Optics*, 17(22):3562–3570, 1978.

[41] El E Fenimore and TM Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[42] Yossi Gandelsman, Assaf Shocher, and Michal Irani. "Double-DIP": Unsupervised image decomposition via coupled deep-image-priors. In *Computer Vision and Pattern Recognition (CVPR), 2019 IEEE Conference on*, 2019.

[43] Genevieve Gariepy, Francesco Tonolini, Robert Henderson, Jonathan Leach, and Daniele Faccio. Detection and tracking of moving objects hidden from view. *Nature Photonics*, 10(1):23–26, 2016.

[44] Helmut Gernsheim. *A concise history of photography.* Number 10. Courier Corporation, 1986.

[45] Karl Goldberg. Upper bounds for the determinent of a row sto· chastic matrix, 1. *Res. NBS*, 708:157–158, 1966.

[46] Joseph W Goodman. *Introduction to Fourier optics.* Roberts and Company Publishers, 2005.

[47] Stephen R Gottesman and EE Fenimore. New family of binary arrays for coded aperture imaging. *Applied optics*, 28(20):4344–4352, 1989.

[48] Peter T Gough and David W Hawkins. Unified framework for modern synthetic aperture imaging algorithms. *International journal of imaging systems and technology*, 8(4):343–358, 1997.

[49] Jacques Hadamard. Resolution d'une question relative aux determinants. *Bull. des sciences math.*, 2:240–246, 1893.

[50] Marshall Hall. A survey of difference sets. *Proceedings of the American Mathematical Society*, 7(6):975–986, 1956.

[51] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. Gan-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.

[52] Pat Hanrahan and Wolfgang Krueger. Reflection from layered surfaces due to subsurface scattering. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 165–174, 1993.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[54] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations*, 2019.

[55] Berthold KP Horn. Interesting eigenvectors of the Fourier transform. *Transactions of the Royal Society of South Africa*, 65(2):100–106, 2010.

[56] Robert Alexander Houstoun. *A treatise on light.* Longmans, Green and Company, 1915.

[57] Achuta Kadambi, Hang Zhao, Boxin Shi, and Ramesh Raskar. Occluded imaging with time-of-flight sensors. *ACM Transactions on Graphics (ToG)*, 35(2):1–12, 2016.

[58] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using transient imaging, 09 2009.

[59] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 233–240. IEEE, 2011.

[60] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[61] Marco La Manna, Fiona Kine, Eric Breitbach, Jonathan Jackson, Talha Sultan, and Andreas Velten. Error backprojection algorithms for non-line-of-sight imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1615–1626, 2018.

[62] Edmund Y Lam and Joseph W Goodman. A mathematical analysis of the DCT coefficient distributions for images. *IEEE transactions on image processing*, 9(10):1661–1666, 2000.

[63] Martin Laurenzis, Andreas Velten, and Jonathan Klein. Dual-mode optical sensing: three-dimensional imaging and seeing around a corner. *Optical Engineering*, 56(3):031202, 2016.

[64] Emma Lehmer. On residue difference sets. *Canadian Journal of Mathematics*, 5:425–432, 1953.

[65] Anat Levin and Fredo Durand. Linear view synthesis using a dimensionality gap light field prior. In *In Proc. IEEE CVPR*, pages 1–8, 2010.

[66] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70, 2007.

[67] Anat Levin, William T Freeman, and Frédo Durand. Understanding camera trade-offs through a Bayesian analysis of light field projections. In *European Conference on Computer Vision*, pages 88–101. Springer, 2008.

[68] Anat Levin, Yair Weiss, Fredo Durand, and William Freeman. Understanding and evaluating blind deconvolution algorithms. 2009.

[69] Lianlin Li, Fang Li, Tiejun Cui, Yunhua Tan, and Kan Yao. Far-field imaging beyond the diffraction limit using a single radar. *arXiv preprint arXiv:1406.2168*, 2014.

[70] Tomohiro Maeda, Yiqin Wang, Ramesh Raskar, and Achuta Kadambi. Thermal non-line-of-sight imaging. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2019.

[71] A Mayer and J-P Vigneron. Transfer matrices combined with GreenâĂŹs functions for the multiple-scattering simulation of electronic projection imaging. *Physical Review B*, 60(4):2875, 1999.

[72] Ciaran P Moore, Richard J Blaikie, and Matthew D Arnold. An improved transfer-matrix model for optical superlenses. *Optics Express*, 17(16):14260–14269, 2009.

[73] TWJ Moorhead and TD Binnie. Smart CMOS camera for machine vision applications. In *Image Processing and Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465)*, volume 2, pages 865–869. IET, 1999.

[74] Mankal Narasinha Murthy et al. Sampling theory and methods. *Sampling theory and methods.*, 1967.

[75] MG Neubauer and AJ Radcliffe. The maximum determinant of±1 matrices. *Linear algebra and its applications*, 257:289–306, 1997.

[76] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera.* PhD thesis, Stanford University, 2005.

[77] Raymond EAC Paley. On orthogonal matrices. *Journal of Mathematics and Physics*, 12(1-4):311–320, 1933.

[78] Rohit Pandharkar, Andreas Velten, Andrew Bardagjy, Everett Lawson, Moungi Bawendi, and Ramesh Raskar. Estimating motion and size of moving non-line-of-sight objects in cluttered environments. In *CVPR 2011*, pages 265–272. IEEE, 2011.

[79] Seymour V Parter. On the distribution of the singular values of Toeplitz matrices. *Linear Algebra and its Applications*, 80:115–130, 1986.

[80] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[81] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[82] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 795–804. ACM, 2006.

[83] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

[84] George O Reynolds et al. *The New Physical Optics Notebook: Tutorials in Fourier Optics.* ERIC, 1989.

[85] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 860–867. IEEE, 2005.

[86] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472, 2019.

[87] CJR Sheppard and HJ Matthews. Imaging in high-aperture optical systems. *JOSA A*, 4(8):1354–1360, 1987.

[88] Dongeek Shin, Ahmed Kirmani, Vivek K Goyal, and Jeffrey H Shapiro. Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors. *IEEE Transactions on Computational Imaging*, 1(2):112–125, 2015.

[89] Shikhar Shrestha, Felix Heide, Wolfgang Heidrich, and Gordon Wetzstein. Computational imaging with multi-camera time-of-flight systems. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

[90] James Singer. A theorem in finite projective geometry and some applications to number theory. *Transactions of the American Mathematical Society*, 43(3):377–385, 1938.

[91] Cyril Soler and François X Sillion. Fast calculation of soft shadow textures using convolution. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 321–332, 1998.

[92] Ralph G Stanton and DA Sprott. A family of difference sets. *Canadian Journal of Mathematics*, 10:73–77, 1958.

[93] Seán M Stewart. Spectral peaks and Wien's displacement law. *Journal of Thermophysics and Heat Transfer*, 26(4):689–692, 2012.

[94] Takashi Takenaka, Mitsuhiro Yokota, and Otozo Fukumitsu. Propagation of light beams beyond the paraxial approximation. *JOSA A*, 2(6):826–829, 1985.

[95] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[96] Christos Thrampoulidis, Gal Shulkind, Feihu Xu, William T Freeman, Jeffrey Shapiro, Antonio Torralba, Franco Wong, and Gregory Wornell. Exploiting occlusion in non-line-of-sight active imaging. *IEEE Transactions on Computational Imaging*, 2018.

[97] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 374–381. IEEE, 2012.

[98] Chia-Yin Tsai, Kiriakos N Kutulakos, Srinivasa G Narasimhan, and Aswin C Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[99] Chia-Yin Tsai, Aswin C Sankaranarayanan, and Ioannis Gkioulekas. Beyond volumetric albedo–a surface optimization framework for non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1545–1555, 2019.

[100] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[101] David Van Veen, Ajil Jalal, Eric Price, Sriram Vishwanath, and Alexandros G Dimakis. Compressed sensing with deep image prior and learned regularization. *arXiv preprint arXiv:1806.06438*, 2018.

[102] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Transactions on Graphics (TOG)*, 26(3):69, 2007.

[103] Suresh Venkatesh, Naren Viswanathan, and David Schurig. W-band sparse synthetic aperture for computational imaging. *Optics express*, 24(8):8317–8331, 2016.

[104] Albert Leon Whiteman et al. A family of difference sets. *Illinois Journal of Mathematics*, 6(1):107–121, 1962.

[105] Matthias M Wloka and Brian G Anderson. Resolving occlusion in augmented reality. In *Proceedings of the 1995 symposium on Interactive 3D graphics*, pages 5–12, 1995.

[106] L. Xia, C.C. Chen, and J.K. Aggarwal. Human detection using depth information by Kinect. *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011.

[107] Feihu Xu, Dongeek Shin, Dheera Venkatraman, Rudi Lussana, Federica Villa, Franco Zappa, Vivek K Goyal, Franco Wong, and Jeffrey Shapiro. Photon-efficient computational imaging with a single-photon camera. 2016.

[108] Yongyi Yang, Nikolas P Galatsanos, and Henry Stark. Projection-based blind deconvolution. *JOSA A*, 11(9):2401–2409, 1994.

[109] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell. Analysis and optimization of aperture design in computational imaging. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4029–4033. IEEE, 2018.

[110] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12231–12239, 2019.

[111] M. Young. Pinhole. *Applied Optics*, 10:2763–2767, 1971.

[112] Hansong Zhang, Dinesh Manocha, Tom Hudson, and Kenneth E Hoff III. Visibility culling using hierarchical occlusion maps. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 77–88, 1997.

[113] Li Zhang, Richard C Lanza, Berthold KP Horn, and Robert E Zimmerman. Three-dimensional coded-aperture techniques in diagnostic nuclear medicine imaging. In *Medical Imaging 1998: Physics of Medical Imaging*, volume 3336, pages 364–373. International Society for Optics and Photonics, 1998.

[114] Changyin Zhou, Stephen Lin, and Shree K Nayar. Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72, 2011.