# Variable-Aperture Photography

by

Samuel William Hasinoff

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

# Abstract

Variable-Aperture Photography

Samuel William Hasinoff
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2008

While modern digital cameras incorporate sophisticated engineering, in terms of their core functionality, cameras have changed remarkably little in more than a hundred years. In particular, from a given viewpoint, conventional photography essentially remains limited to manipulating a basic set of controls: exposure time, focus setting, and aperture setting.

In this dissertation we present three new methods in this domain, each based on capturing multiple photos with different camera settings. In each case, we show how defocus can be exploited to achieve different goals, extending what is possible with conventional photography. These methods are closely connected, in that all rely on analyzing changes in aperture.

First, we present a 3D reconstruction method especially suited for scenes with high geometric complexity, for which obtaining a detailed model is difficult using previous approaches. We show that by controlling both the focus and aperture setting, it is possible compute depth for each pixel independently. To achieve this, we introduce the "confocal constancy" property, which states that as aperture setting varies, the pixel intensity of an in-focus scene point will vary in a scene-independent way that can be predicted by prior calibration.

Second, we describe a method for synthesizing photos with adjusted camera settings in post-capture, to achieve changes in exposure, focus setting, *etc.* from very few input photos. To do this, we capture photos with varying aperture and other settings fixed, then recover the underlying scene representation best reproducing the input. The key to the approach is our layered formulation, which handles occlusion effects but is tractable to invert. This method works with the built-in "aperture bracketing" mode found on most digital cameras.

Finally, we develop a "light-efficient" method for capturing an in-focus photograph in the shortest time, or with the highest quality for a given time budget. While the standard approach involves reducing the aperture until the desired region is in-focus, we show that by "spanning" the region with multiple large-aperture photos, we can reduce the total capture time and generate the in-focus photo synthetically. Beyond more efficient capture, our method provides 3D shape at no additional cost.

# Acknowledgements

I am deeply grateful to Kyros Kutulakos for his thoughtful supervision over more than half a decade, and two continents. I have been inspired by his push toward fundamental problems. Our discussions over the years have consistently been challenging and creative, and the most interesting ideas in this dissertation were born from this interaction. I credit him with teaching me how to think about research.

Many thanks are also due to my committee members, Allan Jepson, David Fleet, and Aaron Hertzmann, for their criticism and support throughout this process, and for their specific help improving this document. A special thanks to Sven Dickinson, for his guidance when I first became interested in computer vision, and for his continued encouragement. The department is fortunate to have assembled this group of faculty.

I would also like to acknowledge an inspiring group of colleagues—Anand Agarawala, Mike Daum, Ady Ecker, Francisco Estrada, Fernando Flores-Mangas, Midori Hyndman, Stratis Ioannidis, Nigel Morris, Makoto Okabe, Faisal Qureshi, Daniela Rosu, Jack Wang, and too many others to list individually. Our friendships and shared experiences have defined my life as a graduate student, besides teaching me what to order on Spadina (and Chengfu Lu). Extra thanks to Midori and Jack for their permission to be immortalized as datasets.

This work was supported by the Natural Sciences and Engineering Research Council of Canada under the CGS-D and RGPIN programs, by a fellowship from the Alfred P. Sloan Foundation, by an Ontario Premier's Research Excellence Award, and by Microsoft Research Asia. Chapters 3 and 4 are based on previously published material, reproduced with kind permission of Springer Science+Business Media [53] and IEEE [54].

I reserve the last word of thanks for my family—for their unconditional love, and for supporting my entry into the family business.

# Contents

# Chapter 1

# Introduction

A photograph is a secret about a secret. The more it tells you the less you know.

Diane Arbus (1923–1971)

I believe in equality for everyone, except reporters and photographers.

Mahatma Gandhi (1869–1948)

At the basic optical and functional level, the cameras we use today are very similar to the cameras from more than a hundred years ago. The most significant change has been the tight integration of computation in all aspects of photography, from color processing at the sensor level, to the automatic control of all camera parameters, to the integration of face-detection algorithms to ensure that the subject is focused and well-exposed.

Though modern cameras offer advanced features that can be of assistance to the photographer, all these features are in support of a fundamental question that hasn't changed since the early days of photography—what camera settings should I use?

As experienced photographers know, conventional cameras of all varieties share the same set of basic controls, accessible in "manual" mode: exposure time, focus setting, and aperture setting. So for a given framing of the scene, provided by the viewpoint and zoom setting, our choices for photography are effectively limited to manipulating just three controls. Thus, we can regard any photograph as a point lying in the 3D space defined by the controls for conventional photography (Fig. 1.1).

This model has begun to change in recent years, with the development of new prototype camera designs that try to extend the capabilities of traditional photography. These new designs rely on various strategies such as using ensembles of multiple cameras [48, 61, 74, 77], trading

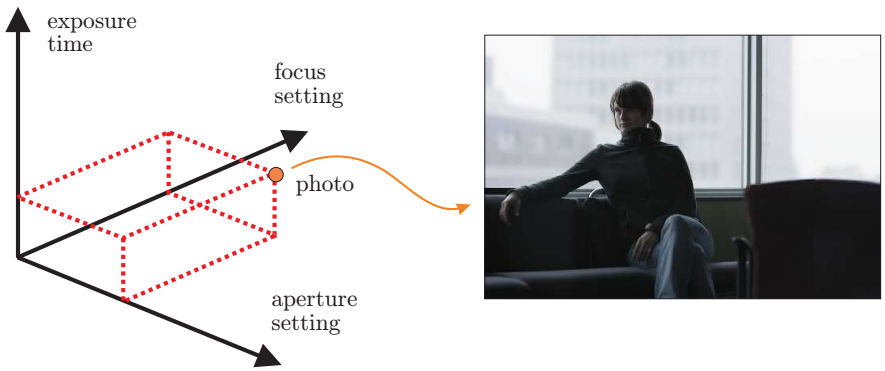



**Figure 1.1:** Basic controls for conventional photography. For a given viewpoint and zoom setting, every photograph we capture with our camera can be thought of as a point in the 3D space of camera settings.

sensor resolution for measurements along new dimensions [9, 47, 49, 73, 75, 85, 115], introducing coded patterns into the optics [32, 36, 58, 69, 96, 103, 115], changing the optics themselves [8, 28, 33, 44, 138], and controlling the environment with projected illumination [34, 79, 93].

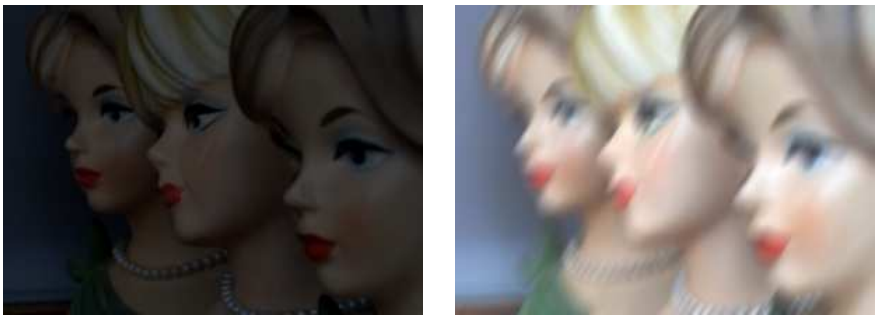
While some of this recent work is very exciting, in this dissertation we revisit the domain of conventional photography, and advocate taking multiple photographs from a fixed viewpoint. We will see that when camera settings are chosen appropriately, conventional photographs can reveal deep structure about the scene, and that limitations usually attributed to the conventional camera design can be overcome. An obvious advantage of restricting ourselves to standard cameras is that the methods we propose can be put into immediate practice.

Despite its apparent simplicity, we will demonstrate how conventional photography can be used to address a wide variety of fundamental questions:

- How can we resolve fine 3D structure for scenes with complex geometry?
- How can we capture a small number of photos that enable us to manipulate camera parameters synthetically, after the capture session?
- How do we capture an in-focus and well-exposed photo of a subject in the least amount of time?
- How do we capture the best possible photo of a subject within a restricted time budget?

Beyond our specific contributions in these areas, the insights we develop can be applied more broadly, and guide the development of general camera designs as well.

Our work follows a well-established strategy in computer vision of capturing multiple photos from the same viewpoint with different camera settings. Most of the work along these lines has concentrated on combining differently focused images, for the purpose of computing depth [30, 43, 60, 64, 80, 92, 111, 120, 129], or for restoring the underlying in-focus scene [10, 130]. We



(a) (b)

**Figure 1.2:** Varying the exposure time. (a) A short exposure time (0.1 s) leads to a dark, relatively noisy image. (b) With a longer exposure time (0.4 s), the image is brighter, but suffers from blur due to the motion of the camera over the exposure interval. The same scene is shown in Fig. 1.5 without motion blur. © dpreview.com

review this work in Chapter 2. Other methods in this domain have explored varying exposure time to increase the dynamic range [31, 78], and capturing multiple reduced-exposure images to address motion blur [20, 113, 131]. We discuss these related methods later, as they relate to our specific approaches.

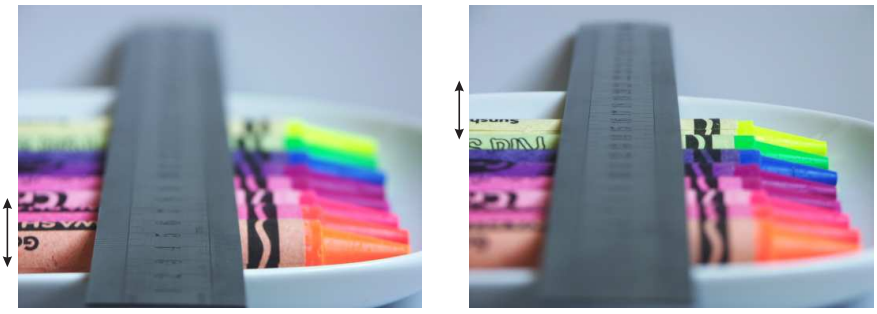
Collectively, we refer to our work as *variable-aperture photography*, because a connecting feature is that all of our methods rely on analyzing changes in aperture—a camera control that hasn't received much attention until lately. Our methods are based on three basic principles: taking multiple photographs with different camera settings, analyzing properties of defocus in detail, and maximizing the light-gathering ability of the camera.

## 1.1 Controls for Photography

To put our methods in context, we first give a high-level overview the basic camera controls (Fig. 1.1) and the tradeoffs that they impose. This discussion should be familiar to technically-minded photographers, who face the practical challenge of manipulating these controls to capture compelling images of their subjects.

**Exposure Time.** The simplest camera control is exposure time, which determines the amount of time that the shutter remains open and admits light into the camera. Clearly, longer exposure times allow the sensor to collect more light, leading to brighter images (Fig. 1.2). Exposure time does not depend on any mechanical lens adjustments, nor even the presence of a lens.

The advantage of brighter images is that up to the saturation limit of the sensor, brighter pixels have relatively lower noise levels [56, 76]. In practice, the exposure time must be chosen



(a) (b)

**Figure 1.3:** Varying focus setting, from (a) closer to the camera, to (b) further from the camera. The intervals illustrate the depth of field, which can also be observed along the ruler. The further away that a region of the scene lies from the depth of field, the more detail is lost due to defocus. © dpreview.com

carefully to avoid excessive saturation, because wherever the subject is over-exposed, all information is lost except for a lower bound on brightness.

While exposure time is a key mechanism for selecting image brightness, it also presents us with an important tradeoff. In particular, the long exposures necessary to obtain bright images open the possibility that the image will be degraded due to motion over the capture (Fig. 1.2b). Both subject motion and motion of the camera are potential sources of this blurring, so all else being equal we would prefer to keep the shutter open as briefly as possible [20, 113, 131].

**Focus Setting.** While the focus setting does not affect brightness, it controls the distance in the scene at which the scene appears at its sharpest (Fig. 1.3). In contrast to the idealized pinhole model, in which every pixel on the sensor plane corresponds to a single ray of light (Sec. 2.2.1), integration over the lens means that only light from certain 3D points in the scene will be perfectly focused to the sensor plane.

In theory, each focus setting defines a unique distance from which scene points are brought into perfect focus. In practice, however, there is a whole range of distances known as the *depth of field* (DOF) for which the degree of blur is negligible. On a practical level, if we want the subject to be sharp, it must lie within the depth of field. The further away we get from the depth of field, the more detail will be lost due to defocus.

Note that since focus setting is controlled by modifying the effective lens-to-sensor distance, it typically has the side-effect of magnifying the image and producing more subtle geometric distortions as well (Sec. 2.2.3).

f/16 (3.1 mm) f/5.6 (8.9 mm) f/1.8 (27.8 mm)

**Figure 1.4:** Non-circular variation in aperture for a real 50 mm SLR lens.

(a) (b)

**Figure 1.5:** Varying the aperture setting in "aperture-priority" mode, which adjusts the exposure time to keep image brightness roughly constant. (a) A small aperture (f/8) yields a large depth of field, with most of the scene "acceptably" in focus, whereas (b) a larger aperture (f/1.4) yields a shallower depth of field, with a more restricted region of the scene in focus. © dpreview.com

**Aperture Setting** The final and most complex camera control is aperture setting, which affects the diameter of a variable-size opening in the lens (Fig. 1.4) that lets light enter the camera. Changing aperture is particularly interesting because it has two interconnected effects. First, larger apertures collect more light, so in the same exposure time, photos taken with a larger aperture will be more brightly exposed. Secondly, larger apertures increase the level of defocus for every point in the scene, leading to a reduction in the depth of field (Fig. 1.5).

The light-gathering ability of large apertures is useful in two ways: it can lead to brighter images with lower noise levels, and it can also lead to faster exposure times for reduced motion blur. The important tradeoff of using wide apertures is that a greater portion of the scene will appear defocused.

By convention, aperture setting is written using the special notation f/$\alpha$. This corresponds to an effective aperture diameter of $f/\alpha$, where $f$ is the focal length of the lens (see Sec. 2.2.1), and $\alpha$ is known as the f-number of the lens. In modern auto-focus lenses, the aperture is usually discretized so that its effective area doubles with every three discrete steps.

Changing the aperture setting also leads to secondary radiometric effects, most notably in-

creased vignetting, or relative darkening at the corners of the image (Sec. 2.2.3).

**Space of Conventional Photographs.** In summary, by manipulating the three basic camera controls, we can capture photographs that vary in terms of their brightness level and defocus characteristics. Motion blur will be more severe with longer exposure times—if motion is constant over the exposure interval, its magnitude will be roughly proportional to exposure time.

Both image brightness and defocus depend on the interaction of two camera controls. On one hand, image brightness is directly related to the combination of exposure time and aperture area. On the other hand, defocus depends on the combination of aperture and focus setting, but in orthogonal ways—the aperture setting controls the extent of the depth of field, whereas the focus setting controls its distance from the camera. Together, the aperture and focus settings fully determine how defocus varies with depth.

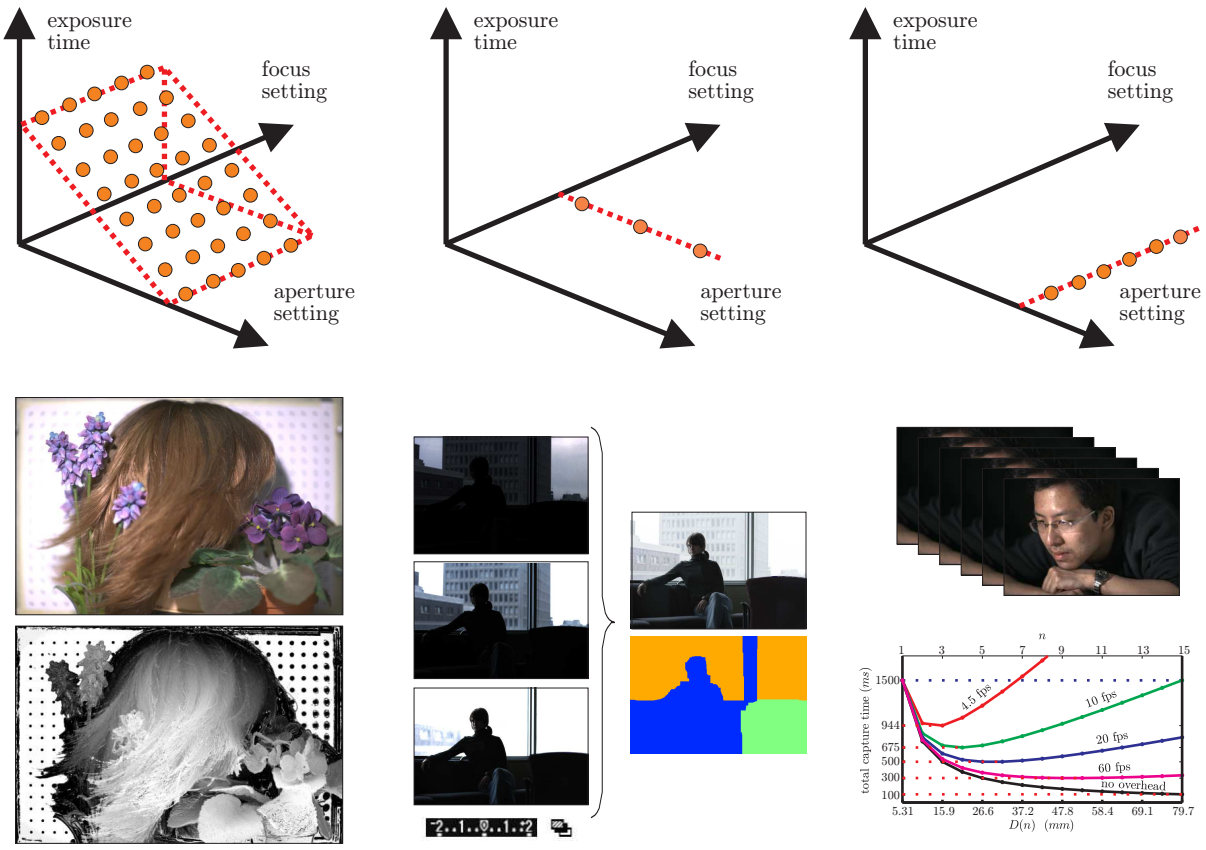
## 1.2 Overview

After presenting background material on the analysis of defocus (Chapter 2), we describe three new methods for *variable-aperture photography* based on applying computation to conventional photographs. Despite our seemingly restrictive domain—manipulating basic camera controls from a fixed viewpoint—we show that identifying the right structure in the space of photographs allows us to achieve gains in 3D reconstruction, in resynthesis, and in efficiency (Fig. 1.6).

**Confocal Stereo.** In our first method, we show that by varying both aperture and focus setting—holding image brightness roughly constant—it is possible compute depth for each pixel independently (Chapter 3). This allows us to reconstruct scenes with very high geometric complexity or fine-scale texture, for which obtaining a detailed model is difficult using existing 3D reconstruction methods.

The key to our approach is a property that we introduce called "confocal constancy". This property states that we can radiometrically calibrate the lens so that under mild assumptions, the color and intensity of an in-focus point projecting to a single pixel will be unchanged as we vary the aperture of the lens.

To exploit this property for reconstruction, we vary the focus setting of the lens and, for each focus setting, capture photos at multiple aperture settings. In practice, these photos must be "aligned" to account for the geometric and radiometric distortions as aperture and focus varies. Because our method is designed for very high-resolution photos, we develop detailed




**Confocal Stereo**
pixel-level 3D shape
for complex scenes
(Chapter 3)

**Multiple-Aperture Photography**
refocusing in
high dynamic range
(Chapter 4)

**Light-Efficient Photography**
faster, more
efficient capture
(Chapters 5-6)

**Figure 1.6:** High-level overview. This dissertation explores what can be accomplished by manipulating basic camera controls (Fig. 1.1) and combining multiple photos from the same viewpoint. We develop new methods in this domain that contribute to three different areas: capturing highly detailed 3D geometry, enabling post-capture resynthesis, and reducing the time required to capture an in-focus photo.

calibration methods to achieve this alignment.

The other important idea of our approach is that we can organize the aligned photos into a set of aperture-focus images (AFIs), one for each pixel, that describe how an individual pixel's appearance varies across aperture and focus. In this representation, computing the depth of a pixel is reduced to processing its AFI to find the focus setting most consistent with confocal constancy.

Together, these ideas lead to an algorithm we call *confocal stereo* that computes depth for each pixel's AFI independently. This lets us reconstruct scenes of high geometric complexity, with more detail than existing defocus-based methods.

**Multiple-Aperture Photography.** The next method we describe lies at the other end of the spectrum in terms of the number of input photos required. We show that by capturing several photos with varying aperture and keeping other settings fixed, we can recover a scene representation with increased dynamic range that also allows us to synthesize new photos with adjusted camera settings (Chapter 4). This method greatly increases the photographer's flexibility, since decisions about exposure, focus setting, and depth of field can be deferred until after the capture session.

Our method, that we call *multiple-aperture photography*, can be thought of as an extension of standard high dynamic range photography [31, 78], since it uses the same number of input photos with similar variation in image brightness. As we show, by analyzing defocus across photos with different aperture settings, not only can we recover the in-focus high dynamic range image, but also an approximate depth map. It is this richer representation of in-focus radiance plus depth that lets us synthesize new images with modified camera settings.

The key to the success of our approach is the layered formulation we propose, which handles defocus at occlusion boundaries, but is computationally efficient to invert. This model lets us accurately account for the input images, even at depth discontinuities, and makes it tractable to recover an underlying scene representation that simultaneously accounts for brightness, defocus, and noise.

On a practical level, another benefit of this method is that we can capture the input photos by taking advantage of the one-button "aperture bracketing" feature found on many digital cameras.

**Light-Efficient Photography.** The last method we describe addresses the question of how we capture an in-focus and well-exposed photograph in the shortest time possible (Chapter 5). We show that by "spanning" the desired depth of field with multiple large-aperture photos, we can reduce the total capture time compared to the basic single-photo approach, without sacrificing image noise. Under this approach, we generate the desired in-focus photo synthetically, by applying compositing techniques to our input. Beyond more efficient capture, this has the important benefit of providing approximate 3D shape at no additional acquisition cost.

This method, which we call *light-efficient photography*, starts from a simple observation that large apertures are generally more efficient than small ones, in the sense that their increased light-gathering ability more than compensates for their reduced depth of field. We formalize this idea for lenses both with continuously-variable apertures and with discrete apertures, with all photos captured at the same optimal exposure level. Our analysis provides us with the provably time-optimal capture sequence spanning a given depth of field, for a given level of camera

overhead.

In a recent extension to this work, we have also analyzed the related problem of capturing the highest-quality in-focus photo when we are constrained to a time budget (Chapter 6). The analysis in this case is more complex, since we can no longer assume that exposure level is fixed at the optimal level, therefore we must also consider tradeoffs between image noise and defocus. To do this in a principled way, we propose a detailed imaging model that allows us to predict the expected reconstruction error for a given capture strategy. Our preliminary results show that the previous solution, spanning the depth of field with wide-aperture photos, is generally optimal in these terms as well, provided that the time budget is not overly constrained (*i.e.*, that we have on the order of 1/300-th or more of the previous time budget). For severely constrained time budgets, it is more beneficial to span the depth of field incompletely and accept some defocus in expectation.

# Chapter 2

# Lenses and Defocus

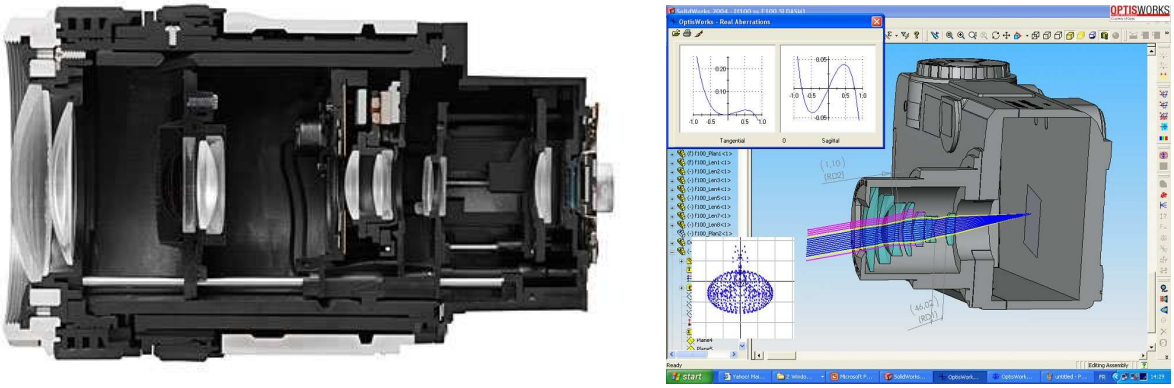
> You cannot depend on your eyes when your imagination is out of focus.
>
> *A Connecticut Yankee in King Arthur's Court*
> Mark Twain (1835–1910)

In classical optics, the convergence of light to a sharp point, or *focus*, has been studied for hundreds of years [105]. While photographers and artists commonly use defocus for expressive effect, in image analysis, defocus is typically regarded as a form of degradation, corrupting the ideal pinhole image. Indeed, the fine image detail lost to defocus cannot be recovered in general, without prior knowledge about the underlying scene.

Although defocus can be thought of as a form of degradation, it also has a particular structure that encodes information about the scene *not* present in an ideal pinhole image. In particular, the depth of a given point in the scene is related to its degree of defocus.

Using a stationary camera with varying settings to measure defocus is particularly well-suited to reconstructing scenes with large appearance variation over viewpoint. Practical examples include scenes that are highly specular (crystals), finely detailed (steel wool), or possess complex self-occlusion relationships (tangled hair). For this reason, 3D reconstruction methods from defocused images hold great potential for common scenes for which obtaining detailed models may be beyond the state of the art [57, 132, 133, 136, 137]. A further advantage of reconstruction methods using defocus is the ability to detect camouflaged objects, which allows segmentation of the scene based on shape rather than texture cues [39].

In this chapter, we review models for lenses and defocus used in computer vision, and survey previous defocus-based methods used for 3D reconstruction [30, 43, 60, 64, 80, 92, 111, 120, 129], or for synthesizing new images from the recovered underlying scene [10, 54, 87, 130, 135].



(a) (b)

**Figure 2.1:** (a) Cut-away view of a real zoom lens, the Panasonic Lumix DMC-FZ30, reveals 14 lens elements arranged in 10 groups. In response to user-applied focus and zoom settings, some of the lens groups translate along the optical axis by different amounts. © Panasonic. (b) Screenshot of Optis SolidWorks computer-aided design software [1], which allows lens designs to be modeled using physically-based ray-tracing. © OPTIS.

## 2.1 Parameters for Real Lenses

The minimalist design for a photographic lens consists of a single refractive lens element, at a controllable distance from the sensor plane, with a controllable aperture in front [105]. By contrast, modern commercially available SLR lenses are significantly more complex devices, designed to balance a variety of distortions (Sec. 2.2.2) throughout their range of operation.

Modern lenses are typically composed of 5 or more lens elements, and up to 25 elements is not uncommon for a telephoto zoom lens [2] (Fig. 2.1a). In practice, these lens elements are arranged in fixed groups, whose axial spacing controls the behavior of the lens. Compared to zoom lenses, fixed focal length or *prime* lenses require fewer elements.

The most common lens element shape is the spherical segment, because of its first-order ideal focusing property [105], and the ease with which it can be machined precisely. Modern lens designs often include several aspheric, or non-spherical, lens elements as well, which provide greater flexibility but are more demanding to manufacture.

Despite their complexity, modern SLR lenses are controlled using a set of three basic parameters. We already described two of these parameters, focus setting and aperture setting, in Sec. 1.1. The remaining lens parameter, zoom setting, is only applicable to so-called zoom lenses.[1] Note that from the photographer's point of view, these basic lens parameters are the

[1]More specialized lens designs, such as tilt-shift lenses, offer additional controls, but such lenses are outside the scope of this work.

(a) (b)

**Figure 2.2:** Changing the zoom setting from (a) telephoto (100 mm), to (b) wide-angle (28 mm), but moving the camera to keep the figurine roughly constant-size in the image. With this compensation the depth of field remains nearly constant, despite popular belief to the contrary. Note how flattened perspective in the telephoto case causes the background to be relatively magnified. © dpreview.com

only way to control how the scene is focused onto the image plane. We will discuss these lens parameters more concretely in Sec. 2.2.1, by describing analytic lens models explicitly defined in terms of these parameters.

**Zoom lenses.** While zoom setting is typically held fixed in the context of analyzing defocus, we describe its effects for completeness. On a zoom lens, the zoom setting controls the effective focal length, which in turn determines its focusing power, or degree to which incoming light is redirected.

The main effect of changing the zoom setting is to change the field of view and magnification. Another notable side-effect of changing the zoom setting is the apparent perspective distortion when the subject is kept at a constant size in the frame (Fig. 2.2). Large (telephoto) focal lengths correspond to a narrow field of view, high magnification, and apparently flattened depth variation. By contrast, small (wide angle) focal lengths correspond to a wide field of view, low magnification, and apparently exaggerated depth variation.

The zoom setting has a subtle effect on the focusing behavior of a lens [17]. While telephoto lenses appear to create a shallower depth of field, this effect is primarily due to magnification and perspective flattening, which causes the background to be relatively magnified without resolving any additional detail. If we compensate for the magnification of a given subject, *i.e.*, by moving the camera and adjusting the focus accordingly, the depth of field remains nearly constant across focal length (Fig. 2.2), however slight differences still remain.

**Mechanical implementation.** Most lens designs realize the three lens parameters in standard ways. For example, the lens aperture is usually formed using a set of 5–12 opaque mechan-

ical blades that pinwheel around to block off the opening (Fig. 1.4). While arbitrary aperture masks can be implemented in theory, *e.g.*, using specially designed filters [36], standard cameras use a nested set of approximately circular openings. Note that lenses effectively have internal apertures that block the incoming light as well (see Sec. 2.2.2), but these apertures are not directly controllable.

Changes to the focus setting can be realized by translating the entire assembly of lens elements together, in an axial direction perpendicular to the sensor plane. In practice, changing the focus setting adjusts the inter-element spacing as well, to compensate for distortions. Note that the minimum focusing distance is limited by engineering constraints such as the maximum physical extension of the lens.

Similarly, changes to the zoom setting are effected by modifying the relative spacing of various groups of lens elements. Because zoom lenses must compensate for distortions over wider ranges of focal lengths, they require more lens elements and are more mechanically complex than fixed focal length lenses.

## 2.2 Lens Models and Calibration

Any particular lens model needs to specify two things: (1) geometric properties, or how incoming light is redirected, and (2) radiometric properties, or how light from different incoming rays is blocked or attenuated. Specifying such a model completely defines the image formation process, and allows us to apply the lens model synthetically to a particular description of the scene.

In practice, simple analytic models (Sec. 2.2.1) are typically used to approximate the behavior of the lens, after factoring out the effects of various geometric and radiometric distortions (Sec. 2.2.2). The parameters of these analytic models may be provided by the lens manufacturer, but more often are fit empirically using a calibration procedure (Sec. 2.2.3).

The most detailed lens models available consist of physical simulations of the optics, given a complete description of the lens design [1, 84] (Fig. 2.1b). Unfortunately, designs for commercially available SLR lenses are proprietary, and are not provided with sufficient detail to be used for this purpose. Therefore, to achieve a high level of accuracy one typically must resort to empirical models based on calibration (Sec. 2.2.3). In practice, such empirical models may be valuable for describing individual lenses, which may not be manufactured exactly to specification.

By contrast, some methods such as depth-from-focus (Sec. 2.6) require no explicit lens model whatsoever. These methods instead exploit generic lens properties such as perfect fo-

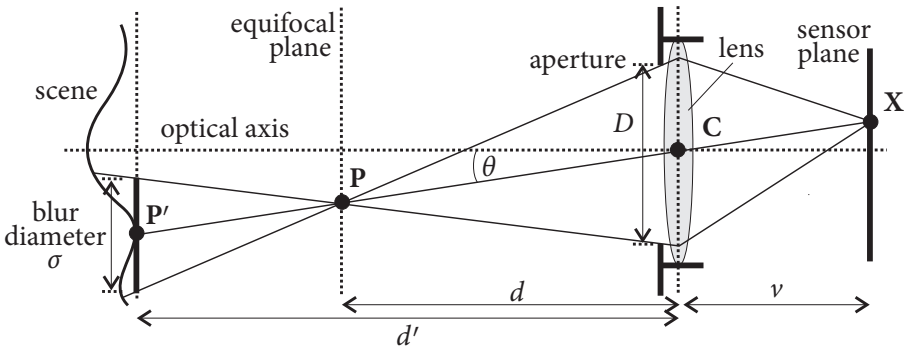



**Figure 2.3:** Geometry of the thin lens model, represented in 2D. When the sensor plane is positioned at an axial distance of $v$ from the lens, the set of in-focus points lie on a corresponding equifocal plane at an axial distance of $d$, as given by Eq. (2.1). As shown, the rays converging on a particular point on the sensor plane, **X**, originate from a corresponding in-focus scene point, **P**. When the scene surface and **P** do not coincide, **X** is defocused and integrates light from a cone, whose projected diameter, $\sigma$, is given by Eq. (2.4). Note that the principal ray passing through the lens center, **C**, is undeflected. The aperture has diameter $D = f/\alpha$, where $f$ is the focal length.

cusing, or "confocal constancy" in the case of confocal stereo (Chapter 3).

### 2.2.1 Basic analytic models

**Pinhole model.** The simplest lens model available is the pinhole model, representing an idealized perspective camera where everything is in-focus. In practice, very small apertures such as f/22 can approximate the pinhole model, however diffraction limits the sharpness that can be achieved with small apertures [105]. Another limitation of small apertures is that they gather less light, meaning that they require long exposure times or strong external lighting.

The pinhole model is specified by its center of projection, **C**, which is coincident with the infinitesimal pinhole aperture (Fig. 2.3). This geometry implies that every point, **X**, on the sensor plane corresponds to a single ray from the scene, $\overrightarrow{\mathbf{PC}}$, therefore the entire image will be in-focus. Photometrically, the image irradiance, $E$, depends only on the radiance, $L$, associated with the corresponding ray. Assuming a linear sensor response, we have $E(\mathbf{X}) \propto L(\overrightarrow{\mathbf{PC}})$.

Note that the pinhole does not redirect light from the scene, but simply restricts which rays reach the sensor. Therefore, an alternate way of thinking about a pinhole lens is as a mechanism to select a 2D image slice from the 4D light field of the scene [48, 74].

Although aperture and zoom setting have no meaningful interpretation for a pinhole, the distance from the pinhole to the sensor plane, $v$, can be interpreted as a degenerate form of focus setting (Fig. 2.3). For any such distance, the pinhole model will still achieve perfect focus, however, moving the sensor plane has the side-effect of magnifying the image.

**Thin lens model.** The thin lens model is a simple, widely used classical model accounting for lenses with variable aperture, focus setting, and focal length (Fig. 2.3). In physical terms, the thin lens model consists of spherical refracting surfaces with negligible separation, and assumes a first-order approximation of geometric optics, where the trigonometric functions are linearized as $\sin(x) \approx x$ and $\cos(x) \approx 1$. For an infinitesimal aperture, the thin lens model reduces to the pinhole model.

The thin lens model is based on a distinguished line known as the *optical axis*. The optical axis is perpendicular to the sensor plane, passes through the lens center, $\mathbf{C}$, and is normal to both refracting surfaces (Fig. 2.3). According to first-order optics, the angle between a ray and the optical axis is negligible. This approximation, also known as the *paraxial* assumption [105], provides invariance to transversal shift perpendicular to the optical axis.

An important consequence of the paraxial assumption is that for a given focus setting, specified by the lens-to-sensor distance, $v$, the surface defining the corresponding set of perfectly focused scene points is a plane parallel to the sensor plane. In other words, the "equifocal" surfaces for the thin lens model are fronto-parallel planes.

Under the paraxial assumption, a spherical refracting surface can be shown to focus incident rays of light to a common point. Then, using basic geometry, we can derive the classic focusing relationship between points on either side of the lens, also known as the *thin lens law*:

$$\frac{1}{v} + \frac{1}{d} = \frac{1}{F} \ , \tag{2.1}$$

where $v$ is the axial distance from a point on the sensor plane to the lens, $d$ is the axial distance from the lens to the corresponding in-focus scene point, and $F$ is the focal length (see Sec. 2.1). Note that under the thin lens model, the focal length also corresponds to the distance behind the lens at which the rays parallel to the optical axis, *i.e.*, from an infinitely distant scene point, will converge.

For a given point on the sensor plane, $\mathbf{X}$, the ray passing through the lens center, $\overrightarrow{\mathbf{XC}}$, also known as the *principal ray*, will not be refracted. This follows from the paraxial assumption, which views the principal ray in the same way as the optical axis. By definition, the corresponding in-focus scene point, $\mathbf{P}$, must lie on the principal ray, giving rise to the following explicit

construction,

$$v = \|\mathbf{C} - \mathbf{X}\| \cos\theta \tag{2.2}$$

$$\mathbf{P} = \mathbf{X} + \frac{d+v}{v}(\mathbf{C} - \mathbf{X}) \ , \tag{2.3}$$

where $\theta$ is the angle between the principal ray and the optical axis (Fig. 2.3), and the scene-side axial distance $d$ may be computed according to Eq. (2.1).

If the closest scene point along the principal ray, $\mathbf{P}'$, lies on the equifocal plane, *i.e.*, if $\mathbf{P}' = \mathbf{P}$, then the corresponding pixel on the sensor plane, $\mathbf{X}$, is perfectly in-focus. Otherwise, $\mathbf{X}$ integrates light from some region of the scene.

From simple geometry, this domain of integration is a cone whose cross-section is the shape of the aperture. At the point where the principal ray meets the scene, we define a "blur circle" describing the extent of defocus, as the intersection of the integration cone with a plane parallel the sensor. By similar triangles, the diameter, $\sigma$, of this blur circle satisfies

$$\sigma = D\frac{|d' - d|}{d} \ , \tag{2.4}$$

where $D = f/\alpha$ is the aperture diameter, and $d'$ is the axial distance of $\mathbf{P}'$ from the lens. By rearranging this equation, we obtain:

$$d' = d\left(1 \pm \frac{\sigma}{D}\right) \ , \tag{2.5}$$

which expresses the depth of the scene, $d'$, in terms of the degree to which pixel $\mathbf{X}$ is defocused, as represented by the blur diameter, $\sigma$. This is the basic idea that enables depth-from-defocus methods (Sec. 2.7).

The image irradiance under the thin lens model depends not only on the radiance associated with the corresponding cone, but also on geometric factors causing falloff over the sensor plane. Assuming a linear sensor response, the thin lens irradiance can be derived as [11, 105]:

$$E(\mathbf{X}) \propto \frac{A\cos^4\theta}{v^2} L(\overrightarrow{\mathbf{PC}}) \ , \tag{2.6}$$

where $A = \frac{1}{4}\pi D^2$ is the area of the aperture, and the radiance from $\mathbf{P}$ is assumed to be constant over the cone of integration. The angle from the optical axis, $\theta$, effectively foreshortens both the aperture and the scene; it also appears in the inverse-squared falloff, which is defined according

to the distance to the sensor plane, $v/\cos\theta$.

Note that the thin lens formula is an idealization, satisfied only for an aberration-free lens near the optical axis, so that the paraxial assumption holds. In practice, the model is still a reasonable approximation for many lenses, however calibrating its parameters is non-trivial (Sec. 2.2.3). For additional accuracy, more detailed empirical calibration may be used to factor out residual geometric and radiometric distortions, to reduce the behavior of a real lens to the thin lens model.

**Thick lens model.** Another classical imaging model favored by some authors is the thick (or Gaussian) lens model [64, 109, 111]. The thick lens model defines two distinct refracting surfaces with fixed separation, where axial distance $d$ and $v$ measured with respect to those planes. However this can easily be reduced to the thin lens model, provided that the medium, *e.g.*, air, is the same on both sides of the lens. In any case, the "thickness" of the lens model has no physical meaning for real multi-element lenses.

**Pupil-centric model.** As Aggarwal and Ahuja note [11], the thin lens model assumes that position of the aperture is coincident with the effective scene-side refractive surface, however real lens designs often violate this assumption. To address this deficiency, they propose a richer analytic model, called the *pupil-centric* model, which incorporates the positions of entrance and exit pupil, and possibly the tilt of the sensor plane relative to the optical axis.

For a given setting of the lens parameters, the pupil-centric model reduces to an instance of the thin lens model, whose effective parameters could be calibrated empirically. The real advantage of the pupil-centric model is that it provides a more accurate analytic model across all lens settings, from a small number of extra model parameters. These pupil-centric parameters may be fit empirically through calibration, though the authors suggest measuring some of them directly, using a second camera to perform depth-from-focus (Sec. 2.6) on the internal components of the lens.

### 2.2.2 Distortions in real lenses

Analytic imaging models, like the thin lens model, serve as a useful first approximation to the behavior of real lenses. In practice, however, real lenses suffer from significant geometric and radiometric distortions from those basic models, also known as aberrations. The bulk of these distortions are due to fundamental limitations in the analytic model, *i.e.*, the approximate first-order model of optics assumed by the thin lens model. However, physical design constraints,

such as aperture placement, as well as limited manufacturing tolerances can also contribute to these distortions.

**Seidel aberrations.** The first category of distortions we consider are geometric distortions from the first-order paraxial model of optics, which prevent rays from the scene from focusing perfectly on the sensor, or from focusing at the expected location. Five common types of geometric distortions, known as Seidel aberrations, may be accounted for by considering a richer third-order model of optics [105]:

- **Spherical aberration** A spherical lens is not the ideal shape for focusing, since rays at the margins of the lens are refracted to relatively closer positions, preventing all rays from converging perfectly at a point on the sensor plane.
- **Coma** For large apertures, off-axis scene points will be defocused in a characteristic comet shape, whose scale increases with the angle from the optical axis.
- **Astigmatism** From the point of view of an off-axis scene point, the lens is effectively tilted with respect to the principal ray. This causes foreshortening and leads to focusing differences in the radial and tangential directions.
- **Field curvature** Even using a perfectly focusing aspheric lens element, the resulting equifocal surfaces in the scene may be slightly curved. This incompatibility between the curved shape of the equifocal surfaces and the planar sensor causes fronto-planar objects to be radially defocused.
- **Radial distortion** If the aperture is displaced from the front of the lens, rays through center of the aperture will be refracted, leading to radially symmetric magnification which depends on the angle of the incoming ray, giving straight lines in the scene the appearance of being curved.

Algebraically, third-order optics involves adding an extra Taylor series term to the trigonometric functions, to obtain $\sin(x) \approx x - \frac{1}{3!}x^3$ and $\cos(x) \approx 1 - \frac{1}{2!}x^2$.

**Chromatic aberrations.** Another fundamental type of distortion stems from the dependence of refractive index on the wavelength of light, according to the same physical principle which causes blue light to be more refracted than red light through a prism [105]. In addition to reducing the overall sharpness of the image, chromatic aberrations can also lead to color fringing artifacts at high-contrast edges.

One component to chromatic aberration is axial, which prevents the lens from focusing simultaneously on different colored rays originating from the same scene point. Chromatic

aberration also has a lateral component, causing off-axis scene points to be focused with magnification that is dependent on their color, leading to prism-like dispersion effects. In practice, systems using multiple lens elements with special spacing or different refractive indexes can largely eliminate both types of chromatic aberration.

**Radiometric distortions.** The last category of distortions we consider are radiometric distortions, which cause intensity variations on the sensor even when the radiance of every ray in the scene is constant. The most common type of radiometric distortion is vignetting, which refers to darkening, or even complete occlusion, at the periphery of the sensor. There are a variety of sources for vignetting:

- **Mechanical vignetting** Some light paths are completely blocked by the main aperture, internal apertures, or external attachments such filters or lens hoods.
- **Natural vignetting** Also known as off-axis illumination, natural vignetting refers to the $\cos^4 \theta$ falloff already accounted for by the thin lens model (Sec. 2.2.1), arising from integration over oblique differential solid angles [105].
- **Optical vignetting** The displacement of the aperture from the front of the lens causes portions of the entrance pupil to become effectively occluded for oblique rays. This type of vignetting leads to characteristic "cat's eye" defocus, corresponding to the shape of the aperture becoming eclipsed toward the edges of the image.

Another radiometric distortion related to optical vignetting, known as pupil aberration, is the nonuniformity of radiometric variation of a scene point across the visible aperture. This effect may be especially pronounced for small asymmetric apertures whose centroid is off-axis [12].

As Kang and Weiss showed, it is possible in principle to recover intrinsic camera calibration by fitting models of vignetting to an image of a diffuse white plane [63]. This demonstrates that radiometric distortions can not only be accounted for, but even carry useful information about the imaging system.

### 2.2.3 Calibration methods

To relate images captured at different lens settings, they must be aligned both geometrically and radiometrically. Any analytic lens model will predict such an alignment, so the simplest approach to camera calibration is to take these parameters directly from the specifications of the lens [111].

For higher accuracy, however, known calibration patterns may be used to estimate the parameters of the analytic lens model empirically [15, 30, 53, 66, 124] (Sec. 3.6). Taking this idea

to an extreme, empirical calibration could theoretically be used to reduce lens calibration to a pixel-level table look-up, with entries for every image coordinate, at every tuple of lens parameters [109].

In the following, we describe geometric and radiometric calibration methods formulated in terms of in-focus 3D points, or else in terms of the principal ray through the camera center. We defer discussing the specific form of defocus, including methods for its empirical calibration, to Sec. 2.3.

**Geometric calibration.** Geometric calibration means that we can locate the projection of the same 3D point in multiple images taken with different settings. Real lenses, however, map 3D points onto the image plane in a non-linear fashion that cannot be predicted by ordinary perspective projection. While the main source of these distortions are changes to the focus or zoom setting, the aperture setting affects this distortion in a subtle way as well, by amplifying certain aberrations which cause small changes to the image magnification.

The single geometric distortion with the largest effect is the linear radial image magnification caused by changes to the focus or zoom setting. Such magnification follows from the thin lens model (Sec. 2.2.1), but for greater accuracy the mapping between lens parameters and magnification must be recovered empirically.

For some reconstruction methods, image magnification is not mentioned explicitly [134], or else is consciously ignored [80, 109, 111]. Since such magnification is about 5 % at the image margins, methods that use very low-resolution images or consider large image patches can ignore these effects. Other reconstruction approaches circumvent the image magnification problem by changing the aperture setting instead [91, 92, 109, 111], by moving the object [82], or by changing the zoom setting to compensate [30, 126].

Another approach for avoiding the image magnification effect is to use image-side *telecentric* optics, designed so that the principal ray always emerges parallel to the optical axis [118–120]. The telecentric lens design has the effect of avoiding magnification with sensor plane motion, and has the added benefit of avoiding the any radiometric falloff due to the position of the sensor plane. Telecentric lens designs are realized by placing an additional aperture at an analytically-derived position, so a tradeoff is their reduced light-gathering ability.

A more direct approach for handling image magnification involves fitting this magnification, either directly [15, 66] or in a prior calibration step [30, 53, 124] (Sec. 3.6), which allows us to warp and resample the input images to some reference lens setting. However, as Willson and Shafer note, simply using the center pixel of the sensor is insufficient for accurately modeling

magnification [127].

Beyond simple image magnification with a displaced center, even richer models of geometric distortion, including radial distortion, have been proposed as well [53, 66, 80, 124] (Sec. 3.6). Kubota, *et al*. proposed a hierarchical registration method, analogous to block-based optical flow, for geometrically aligning defocused images captured at different lens settings [66]. Willson implemented extensive geometric calibration as well, by fitting polynomials in the lens parameters, to a full parameterization of the $3 \times 4$ matrix for perspective projection, inferring the degree of these polynomials automatically [124]. Nair and Stewart suggest correcting for field curvature, by fitting a quadratic surface to a depth map obtained by applying their reconstruction method to a known fronto-planar scene [80].

Another class of geometric distortion, not accounted for in analytic optical models, is non-deterministic distortion, caused by random vibrations, both internal and external to the camera [53, 119, 127], hysteresis of the lens mechanism [127], and slight variations in aperture shape (Sec. 3.6). These effects can be especially significant for high-resolution images, and can even occur when the camera is controlled remotely without any change in settings, and is mounted securely on an optical table. To offset non-deterministic distortions, a first-order translational model can be fit to subpixel shifts [53, 119] (Sec. 3.6). Unlike other geometric distortions, which may be calibrated offline, non-deterministic distortion must be recomputed online, in addition to being accounted for in any offline calibration process.

**Radiometric calibration.** Radiometric calibration means that we can relate the intensity of the same 3D point in multiple images taken with different settings. While the main source of radiometric distortion is changes to the aperture setting, the focus and zoom settings affect this distortion in a more subtle way as well, *e.g*., due to the inverse-squared distance falloff to the sensor plane.

Some reconstruction methods that rely on variable-aperture image comparisons do not mention radiometric distortion explicitly [91, 92]. The most common approach to handling radiometric distortion is simply to normalize a given image region by its mean brightness [109, 111, 118]. This normalization provides some invariance to radiometric distortion, provided that the level of distortion does not vary too much across the image region.

Richer models of radiometric variation may also be fit to images of calibration targets such as a diffuse white plane [53, 54, 63]. One approach is to fit a parametric model of vignetting to each single image, *e.g*., off-axis illumination with a simple linear falloff with radius [63]. By contrast, one can use a more direct approach to obtain an empirical measure of variable-aperture

radiometric variation on a per-pixel level [53, 54] (Secs. 3.5 and 4.7).

Another radiometric distortion that must be accounted for is the camera response function, which maps image irradiance to pixel intensity in a non-linear way [31, 50, 78]. By recovering and inverting this function (Secs. 3.5 and 4.7), we can compare measured image irradiances directly, in agreement with the additive nature of light (Sec. 2.2).

## 2.3 Defocus Models

If the rays incident on the lens from a given 3D scene point do not converge to a unique point on the sensor plane, the scene point is considered to be defocused, and the extent of its defocus can be measured according to the footprint of these rays on the sensor plane. Conversely, a point on the sensor plane is defocused if not all rays that converge to that point originate from a single 3D point lying on the scene surface.

Given a concrete lens model describing how every ray in the scene is redirected and attenuated (Sec. 2.2), defocus will be completely defined. But while the analytic lens models we have described (Sec. 2.2.1) lead directly to simple descriptions of defocus, defocus is often treated separately from other aspects of the imaging model.

Although defocus is overwhelmingly modeled as some form of linear filtering, this approximation cannot accurately represent defocus at sharp occlusion boundaries [16, 42]. The general issue is that linear filtering cannot model the contribution of occluded scene points, because the aperture acts as 2D "baseline" of viewpoints leading to an additive form of self-occlusion [99]. In fact, simulating defocus in its generality requires full knowledge of the light field, which adds significant complexity to the reconstruction problem, even for simple Lambertian scenes. By properly modeling occlusions, more general models of defocus predict such effects as the ability to see "behind" severely defocused foreground objects [42, 54, 77] (Chapter 4).

To date, only a few focus-based reconstruction methods have attempted to accurately model defocus at occluding edges [22, 42, 54, 77]. However, most these methods have been limited by computational inefficiency [42], the assumption that depth discontinuities are due to opaque step edges [22], or the assumption that the scene is composed of two surfaces [22, 42, 77].

For some methods such as depth-from-focus (Sec. 2.6), an explicit model for defocus is not necessary. For these methods, knowing that defocus causes attenuation of high frequencies is enough to identify the lens setting for which focus is optimal. While this approach requires no calibration of defocus, it implicitly assumes that the scene geometry is smooth, otherwise occluding foreground objects could contaminate the estimation.

### 2.3.1 Defocus as linear filtering

In computer vision, the dominant approach to defocus is to model it as a form of linear filtering acting on an ideal in-focus version of the image. This model has the advantage that it allows us to describe an observed defocused image, $\mathbf{I}$, as a simple convolution,

$$\mathbf{I} = \mathbf{B}_\sigma * \hat{\mathbf{I}} \ , \tag{2.7}$$

where $\mathbf{B}_\sigma$ is the blur kernel, or 2D point-spread function, $\sigma$ is a parameter corresponding to the level of defocus, and $\hat{\mathbf{I}}$ is the ideal pinhole image of the scene. We assume that the blur kernel is normalized, $\iint \mathbf{B}_\sigma(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$, implying that radiometric calibration between $\mathbf{I}$ and $\hat{\mathbf{I}}$ has been taken into account. The model of defocus as linear filtering follows from Fourier analysis applied to a fronto-parallel scene [105].

The blur kernel acts as a low-pass filter, so that as the image is defocused, contrast is lost and high frequencies are rapidly attenuated. Although the response of the blur kernel need not decay monotonically for all higher frequencies (*i.e.*, side lobes may exist), in any reasonable physical system, none of its local frequency maxima are as high as the DC response.

To make the identification of blur tractable, we typically require that the blur kernel may be parameterized by a single quantity, $\sigma$. This usually involves the further assumption that the blur kernel $\mathbf{B}_\sigma$ is radially symmetric, and can be parameterized according to its radius of gyration (Sec. 2.5.1).

### 2.3.2 Spatially variant filtering

To relax the assumption that the scene consists of a fronto-parallel plane, we can model the blur parameter as spatially varying, *i.e.*, $\sigma(x, y)$, corresponding to a scene that is only locally fronto-parallel [22, 94, 95]. This results in a more general linear filtering,

$$\mathbf{I}(x, y) = \iint_{s,t} \mathbf{B}_{\sigma(s,t)}(x - s, y - t) \cdot \hat{\mathbf{I}}(s, t)\ \mathrm{d}s\, \mathrm{d}t \ , \tag{2.8}$$

which can be thought of as independently defocusing every pixel in the pinhole image, $\hat{\mathbf{I}}(x, y)$, according to varying levels of blur, and then integrating the results. Note that although this defocusing model is no longer a simple convolution, it is still linear, since every pixel $\mathbf{I}(x, y)$ is a linear function of $\hat{\mathbf{I}}$.

In practice, smoothness priors are often introduced on the spatially variant blur, $\sigma(x, y)$, corresponding to smoothness priors on the scene geometry [94, 95]. These priors help regularize

the recovery of $\hat{\mathbf{I}}(x, y)$ from the image formation model of Eq. (2.8), and balance reconstruction fidelity against discontinuities in depth.

### 2.3.3 Windowed linear filtering

In general, the spatially variant filtering model of Eq. (2.8) means that we can no longer relate a particular observed defocused pixel, $\mathbf{I}(x, y)$, to a single blur parameter, $\sigma$. But provided that $\sigma(x, y)$ is constant within a sufficiently large window centered on $(x, y)$, Eq. (2.8) reduces locally to

$$\mathbf{I}(x, y) \;=\; \left[\mathbf{B}_{\sigma(x,y)} * \hat{\mathbf{I}}\right](x, y) \;. \tag{2.9}$$

This observation motivates the popular sliding window model [15, 29, 30, 35, 40, 43, 80, 81, 91, 92, 109, 111, 118, 120], where for a particular pixel, $(x, y)$, we can express defocusing as filtering within its local window,

$$\mathbf{I} \cdot \mathbf{W}_{(x,y)} \;=\; \left(\mathbf{B}_{\sigma(x,y)} * \hat{\mathbf{I}}\right) \cdot \mathbf{W}_{(x,y)} \;. \tag{2.10}$$

where $\mathbf{W}_{(x,y)}$ represents the windowing function centered at $(x, y)$.

The choice of the window size in this model presents a dilemma. While larger windows may improve the robustness of depth estimation because they provide more data, they are also more likely to violate the assumption that the scene is locally fronto-parallel, and lead to a lower effective resolution. Therefore no single window size for a given scene may lead to both accurate and precise depth estimates.

Note that strictly speaking, the geometric model implied by a sliding window is inconsistent, in the sense that two nearby pixels assigned to different depths contradict each other's assumption that the scene is locally fronto-planar, wherever their windows overlap. Therefore, the windowed model is only a reasonable approximation if the scene is smooth enough so that depth within the sliding window can be locally approximated as fronto-parallel.

A problem caused by analyzing overlapping windows in isolation is that blur from points outside the window may intrude and contaminate the reconstruction [80, 109]. This problem can be partially mitigated using a smooth falloff, such as a Gaussian, for the windowing function [43, 92, 109, 111].

### 2.3.4 Defocus for local tangent planes

To generalize the defocus model beyond spatially variant filtering, we can further relax the assumption that the scene is locally fronto-parallel. In particular, by estimating the normal at each

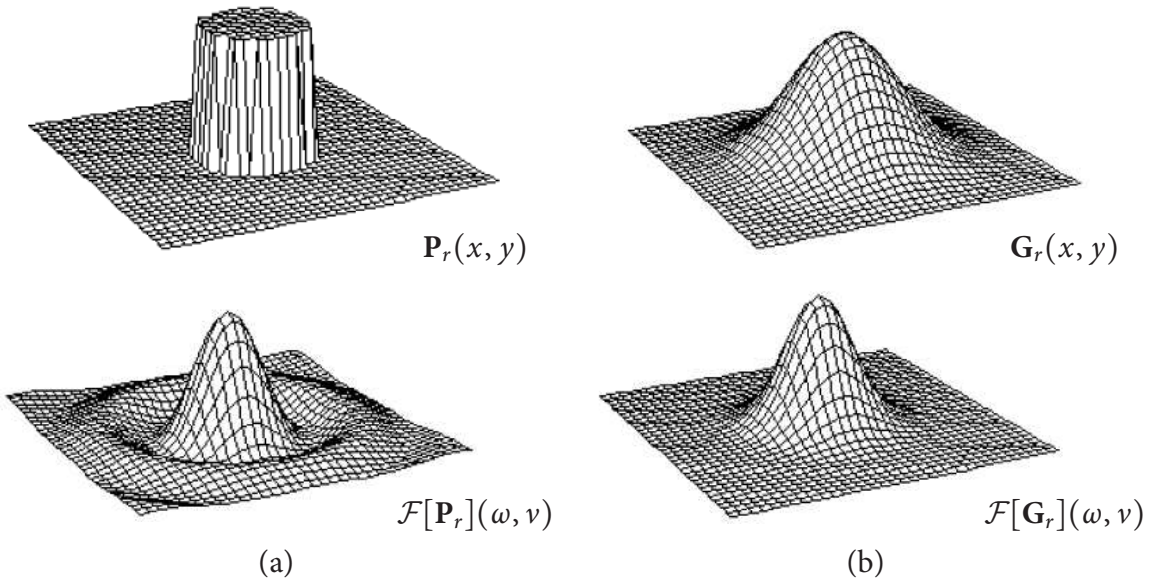



**Figure 2.4:** Point-spread functions for two common defocus models, (a) the pillbox, and (b) the isotropic Gaussian. The top row corresponds to the spatial domain, and the bottom row to the frequency domain.

point as well as its depth, the scene can be modeled as a set of local tangent planes, accounting for the effects of foreshortening on defocus [62, 129]. Note that local tangent planes are not sufficient to model sharp occlusion boundaries or generic self-occluding scenes.

When the surface at a point is locally modeled by a tangent plane, the defocus parameter varies across its neighborhood, meaning that the defocus integral can no longer be expressed using the linear filtering described in Sec. 2.3.2. To address this issue, the defocus integral can be linearized by truncating higher-order terms, assuming that the defocus parameter varies sufficiently smoothly [62, 129].

The local tangent plane model leads to a more complex estimation problem, however it can lead to more stable and accurate estimates compared to the window-based approach, particularly for tilted scenes [62, 129]. Furthermore, the recovered normal is a useful cue for reliability, as severe foreshortening often corresponds to unstable depth estimation.

### 2.3.5 Analytic defocus models

Assuming a linear filtering model of defocus, the two most commonly used analytic models for the blur kernel $\mathbf{B}_\sigma$ are the pillbox and the isotropic Gaussian (Fig. 2.4).

**Pillbox defocus model.** Starting from the thin lens model, geometric optics predict that the footprint of a point on the sensor plane, as projected onto a fronto-parallel plane in the scene,

is just a scaled version of the aperture (Fig. 2.3). So under the idealization that the aperture is circular, the footprint will be circular as well, leading to a cylindrical, or pillbox, model of defocus [99, 120]:

$$\mathbf{P}_r(x,y) = \begin{cases} \frac{1}{\pi r^2} & x^2+y^2 \le r^2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

$$\Longleftrightarrow \quad \mathcal{F}[\mathbf{P}_r](\omega,\nu) = 2\,\frac{J_1\left(2\pi r\sqrt{\omega^2+\nu^2}\right)}{2\pi r\sqrt{\omega^2+\nu^2}}\ , \tag{2.12}$$

where $r = \sigma/2$ is the radius of blur circle (see Fig. 2.3), $\mathcal{F}[\cdot]$ is the Fourier transform operator, and $J_1$ represents the first-order Bessel function, of the first kind, which produces cylindrical harmonics that are qualitatively similar to the 1D function $\mathrm{sinc}(x) = \frac{1}{x}\sin(x)$.

**Gaussian defocus model.** Although first-order geometric optics predict that defocus within the blur circle should be constant, as in the pillbox function, the combined effects of such phenomena as diffraction, lens imperfections, and aberrations mean that a 2D circular Gaussian may be a more accurate model for defocus in practice [43, 92, 109]:

$$\mathbf{G}_r(x,y) = \frac{1}{2\pi r^2}\, e^{-\frac{x^2+y^2}{2r^2}} \tag{2.13}$$

$$\Longleftrightarrow \quad \mathcal{F}[\mathbf{G}_r](\omega,\nu) = e^{-\frac{1}{2}(\omega^2+\nu^2)r^2}\ , \tag{2.14}$$

where $r$ is the standard deviation of the Gaussian. Because the Fourier transform of a Gaussian is simply an unnormalized Gaussian, this model simplifies further analysis. In particular, unlike the pillbox defocus model, the Fourier transform of a Gaussian has no zeros, which makes it more amenable to deconvolution (Sec. 2.7). Under the thin lens model (Fig. 2.3), the blur diameter is proportional to the standard deviation, $\sigma \propto r$.

### 2.3.6 Empirical defocus models

Purely empirical measurements can also be used to recover the blur kernel, with no special assumptions about its form beyond linearity. In blind deconvolution methods (see Sec. 2.7.1), the blur kernel is estimated simultaneously with the geometry and radiance of the perfectly-focused scene.

One common method for calibrating the blur kernel in microscopy applications uses small fluorescent beads mounted on a fronto-planar surface [117], each projecting to approximately

one pixel at the in-focus setting. Since the perfectly focused beads approximate the impulse function, the 2D blur kernel may be recovered directly from the blurred image observed at a given lens setting. By assuming rotational symmetry, the blur kernel may also be recovered empirically from the spread across sharp edges or other known patterns such sine gratings.

Favaro and Soatto suggest an alternative method for recovering defocus calibration, using a more general type of fronto-planar calibration pattern placed at a discretized set of known depths [43]. For each possible scene depth, they propose using a rank-based approximation to recover a linear operator relating defocus between several lens settings, while factoring out the variation due to the underlying radiance.

## 2.4 Image noise models

Having explored various models of how the optics of the lens focus the light from the scene into an image, we briefly review the noisy process by which the sensor measures light [56, 76]. Understanding the sources of noise in the measurement of pixel values can help us estimate the underlying signal more accurately when analyzing defocused images.

**Additive noise.** The most basic model for image noise is additive zero-mean Gaussian noise. Many methods in computer vision assume this generic model, because modeling image noise in a more detailed way is not necessarily helpful—in many practical problems, outliers and modeling errors will dwarf any noise due to the sensor. Additive Gaussian noise follows as a consequence of the central limit theorem, and so it is the appropriate model to use in the absence of any other information. Methods that minimize squared error implicitly assume such a model.

Real image sensors include several sources of noise that can be modeled as additive, namely the noise from the sensor readout, and the noise from the final quantization of the signal [3, 56]. At low exposure levels, these additive noise sources are dominant.

**Multiplicative shot noise.** The basic physical process of detecting photons that arrive at random times corresponds to Poisson-distributed form of noise known as shot noise [56]. Shot noise is multiplicative since the standard deviation of a Poisson-distributed variable is the mean of that variable. In practice, shot noise can be well-approximated using a zero-mean Gaussian noise whose standard deviation is proportional to the raw photon count recorded by the sensor element [76]

For well-exposed photos, shot noise dominates all additive sources of noise. If shot noise is

the only source of noise then signal-to-noise ratio (SNR) will be constant over exposure level; otherwise it will increase for higher exposure levels (see Sec. 5.2).

**Thermal noise.** The final class of image noise comprises thermal effects such as dark current, so-called because this noise will be present even in the absence of any light from the scene. Dark current increases according to the exposure time, and also increases with the temperature of the sensor. Thermal effects for a given sensor are strongest in a particular fixed pattern, which can be mitigated with prior calibration known as dark-frame subtraction [56].

**Transforming the noise.** A variety of transformations are applied to the raw image measurement, both as part of the processing within the camera and later on during image processing (*e.g.*, see Sec. 4.5). These transformations have the important effect of transforming the associated noise as well. A straightforward example is the ISO setting, or gain, which multiplies both the measured signal and its associated noise by a constant factor, before quantization [56]. Another on-camera transformation is the camera response function, which applies an arbitrary monotonic function to the raw image measurement, typically a non-linear gamma-like function function [31, 76]. As a more subtle example, the image processing used for demosaicking the Bayer pattern has the side-effect of introducing spatial correlation to the image noise [76].

## 2.5 Focus Measures

Even in the absence of an explicit model for defocus, it is still possible to formulate a "focus measure" with the ability to distinguish the lens setting at which a given point is optimally in-focus. Such a focus measure is the basis for both image-based auto-focusing [64] and a 3D reconstruction method known as depth-from-focus (Sec. 2.6). We start by analyzing the level of defocus for a known blur kernel, then discuss a variety of possible focus measures for the blind case.

### 2.5.1 Known blur kernel

For a known blur kernel, $\mathbf{B}(x, y)$, a widely used measure of defocus is the radius of gyration [27, 109],

$$\sigma = \left[ \iint (x^2 + y^2)\, \mathbf{B}(x, y)\, \mathrm{d}x\, \mathrm{d}y \right]^{1/2} , \tag{2.15}$$

where $\mathbf{B}$ is assumed to be normalized with zero mean. For an isotropic Gaussian blur kernel, the radius of gyration is equivalent to the standard deviation. Moreover, as Buzzi and Guichard

show, under the assumption that defocus corresponds to convolution, the radius of gyration is the *only* defocus measure that satisfies additivity and several other natural properties [27].

Buzzi and Guichard also remind us that central moments in the spatial domain are related to derivatives at the origin in the frequency domain [27]. This property allows them to reformulate the analytic measure of defocus, Eq. (2.15), in terms of the Laplacian at the DC component in the Fourier domain,

$$\sigma = \left( -\nabla^2 \, \mathcal{F}[\mathbf{B}(x, y)] \right) \Big|_{(0,0)} \quad . \tag{2.16}$$

This implies that defocus can be expressed according to the rate of attenuation at low frequencies, despite the fact that defocus is usually thought of in terms of the extent to which high frequencies are filtered out. To support their argument, Buzzi and Guichard present the results of a small perceptual study, involving images blurred with several artificial defocus functions constructed to preserve high frequencies [27].

### 2.5.2 Blind focus measures

Even when the form of the blur kernel is completely unknown, it may still be possible to detect the lens setting which brings some portion of the scene into optimal focus. To this end, a variety of "blind" focus measures have been proposed, all of which essentially function as contrast detectors within a small spatial window in the image.

The image processing literature is a rich source of ideas for such contrast sensitive filters. One approach is to apply a contrast-detecting filter, such as the image Laplacian [27, 30, 64, 82] (Sec. 2.5.1) or the gradient [64, 126], and to sum the magnitude of those filter responses over the window.

An alternative approach for contrast measurement is to consider the pixel intensities in the patch as an unordered set, ignoring their spatial relationship. Various focus measures along these lines include the raw maximum pixel intensity, the entropy of the binned intensities [64], the kurtosis of the intensities [134], or their variance [64] (Sec. 3.7). Note that by Parseval's theorem, the variance of an image patch is closely related to its total power in the Fourier domain; both give equivalent results when used as a focus measure.

Averaging the focus measure over a patch can cause interference between multiple peaks that represent real structure, so several proposed focus measures explicitly model focus as multimodal [100, 130]. Xu, *et al*. assume a bimodal intensity distribution for the in-focus scene, and define a measure of defocus based on closeness to either of the extreme intensities in the 3D volume consisting of the image window over all focus settings [130]. Their bimodal model of inten-

sity also has the advantage of mitigating bleeding artifacts across sharp intensity edges (Sec. 2.6). Similarly, Schechner, *et al*. propose a voting scheme over the 3D volume, where each pixel votes individually for local maxima across focus setting, and then votes are aggregated over the window, weighted by maxima strength [100].

## 2.6 Depth-from-Focus

Depth-from-focus (DFF) is a straightforward 3D reconstruction method, based on directly applying a blind focus measure (Sec. 2.5) to a set of differently focused photos. For a particular region of the scene, the focus measure determines the focus setting at which the scene is brought into optimal focus, which can then be related to depth, according to prior calibration (Sec. 2.2). DFF has the advantage of being simple to implement and not requiring an explicit calibrated model of defocus (Sec. 2.3).

DFF is most commonly realized by varying the focus setting and holding all other lens settings fixed, which may be thought of as scanning a test surface through the 3D scene volume and evaluating the degree of focus at different depths [64, 80, 126]. Alternative schemes involve moving the object relative to the camera [82].

One disadvantage of DFF is that the scene must remain stationary while a significant number of images are captured with different lens settings. For an online version of DFF, such as image-based auto-focusing, we would prefer to minimize the number of images required. As Krotkov notes, if the focus measure is unimodal and decreases monotonically from its peak, the optimal algorithm for locating this peak is Fibonacci search [64].

Instead of greedily optimizing the focus measure for each pixel independently, it is also possible to construct a prior favoring surface smoothness, and instead to solve a regularized version of DFF, *e.g*., using graph cuts [130].

### 2.6.1 Maximizing depth resolution

To maximize the depth resolution, DFF should use the largest aperture available, corresponding to the narrowest depth of field [99]. This means that a relatively large number of lens settings (up to several dozen) may be required to densely sample the range of depths covered by workspace.

A suggested sampling of depths for DFF is at intervals corresponding to the depth of field, as any denser sampling would mean that the highest frequencies may not be detectably influenced by defocus [99]. Note that the optics predict that depth resolution falls off quadratically with

depth in the scene, according to the quadratic relationship between depth and depth of field [64].

Although the depth resolution of DFF is limited by both the number of images acquired and the depth of field, it is possible to recover depth at sub-interval resolution by interpolating the focus measure about the optimal lens setting, for example, by fitting a Gaussian to the peak [64, 126].

### 2.6.2 Analysis

For DFF to identify an optimal focus peak, there must be enough radiometric variation within the window considered by the focus measure. While an obvious failure case for DFF is an untextured surface, a linear intensity gradient is a failure case for DFF as well, since any symmetric defocus function integrated about a point on the gradient will produce the same intensity [38, 114]. Indeed, theory predicts that for DFF to be discriminative, the in-focus radiance must have non-zero second-order spatial derivatives [37, 38].

Because nearly all blind focus measures (Sec. 2.5.2) are based on spatial image windows, DFF inherits the problems of assuming a windowed, locally fronto-parallel model of the scene (Sec. 2.3.3). A notable exception is the method we present in Chapter 3, which operates at the single-pixel level [53].

Another related problem with DFF is that defocused features from outside the window may contaminate the focus measure and bias the reconstruction to a false peak [80, 109, 130]. This problem may be avoided by considering only image windows at least as large as the largest blur kernel observed over the workspace, but this can severely limit the effective resolution when large blurs are present. Alternatively, Nair and Stewart suggest restricting the DFF computation to a sparse set of pixels corresponding to sufficiently isolated edges [80]. Modeling the intensity distribution as bimodal may also mitigate this problem [130]

## 2.7 Depth-from-Defocus

Depth-from-defocus (DFD) is a 3D reconstruction method based on fitting a model of defocus to images acquired at different lens settings. In particular, the depth of each pixel can be related to its recovered level of defocus, based on the lens model (Sec. 2.2), the defocus model (Sec. 2.3), and the particular lens settings used. In general, DFD requires far less image measurements than DFF, since just two images are sufficient for DFD. Given a strong enough scene model, 3D

reconstruction may even be possible from a single image (Sec. 2.7.3), however DFD methods benefit from more data.

Note that depth recovery using DFD may be ambiguous, since for a particular pixel there are two points, one on either side of the in-focus 3D surface, that give rise to the same level of defocus. For the thin lens model, this effect is represented in Eq. (2.5). In practice the ambiguity may be resolved by combining results from more than two images [111], or by requiring, for example, that the camera is focused on the nearest scene point in one condition [91].

Because nearly all DFD methods are based on linear models of defocus (Sec. 2.3.1), recovering the defocus function can be viewed as a form of inverse filtering, or deconvolution. In particular, DFD is equivalent to recovering $\mathbf{B}_{\sigma(x,y)}$ from Eq. (2.8), or in the simplified case, $\mathbf{B}_{\sigma}$ from Eq. (2.10).

DFD methods can be broken into several broad categories. The most straightforward approach is to tackle the deconvolution problem directly, seeking the scene radiance and defocus parameters best reproducing two or more input images acquired at different camera settings. Alternatively, we can factor out the radiance of the underlying scene by estimating the relative defocus between the input images instead. Finally, if our prior knowledge of the scene is strong enough, we can directly evaluate different defocus hypotheses using as little as a single image.

### 2.7.1 Image restoration

The most direct approach to DFD is to formulate an image restoration problem that seeks the in-focus scene radiance and defocus parameters best reproducing the input images acquired at different lens settings. Note that this optimization is commonly regularized with additional smoothness terms, to address the ill-posedness of deconvolution, to reduce noise, and to enforce prior knowledge of scene smoothness, *e.g.*, [42, 54, 95].

Since a global optimization of the image restoration problem is intractable for images of practical size, such restoration methods resort to various iterative refinement techniques, such as gradient descent flow [62], EM-like alternating minimization [38, 40, 54], or simulated annealing [95]. These iterative methods have the disadvantage of being sensitive to the initial estimate, and may potentially become trapped in local extrema.

**Additive layer decomposition.** One simplifying approach to image restoration is to discretize the scene into additive fronto-parallel depth layers, often one per input image [14, 65, 67, 77]. Unlike layered models incorporating detailed occlusion effects (Sec. 2.3), the layers in this context are modeled as semi-transparent and additive [14, 65, 67, 117]. McGuire, *et al*. suggest

a related formulation where radiance is assigned to two layers, but with an alpha value for the front-most layer represented explicitly as well [77].

This formulation reduces the deconvolution problem to the distribution of scene radiance over the layered volume, where the input images can be reproduced by a linear combination of the layers defocused in a known way. In particular, provided that the input images correspond to an even sampling of the focus setting, the imaging model may be expressed more succinctly as a 3D convolution between the layered scene volume and the 3D point-spread function [75, 117]. This required image restoration can be implemented iteratively, by distributing radiance among the depth layers based on the discrepancy between the input images and the current estimate, synthesized according to the defocus model [14, 65, 67, 75, 77, 117].

This layered additive scene model figures prominently in *deconvolution microscopy* [75, 117], which involves deconvolving a set of microscopy images corresponding to a dense sampling of focus settings, similar to the input for depth-from-focus (Sec. 2.6). Since many microscopy applications involve semi-transparent biological specimens, the assumed additive imaging model is well-justified.

**MRF-based models.** When the layers composing a layered scene model are modeled as opaque instead, every pixel is assigned to a single depth layer, casting depth recovery as a combinatorial assignment problem [10, 54, 95]. This problem can be addressed using a Markov random field (MRF) framework [24], based on formulating costs for assigning discrete defocus labels to each pixel, as well as smoothness costs favoring adjacent pixels with similar defocus labels. Rajagopalan and Chaudhuri formulate a spatially-variant model of defocus (Sec. 2.3.2) in terms of an MRF, and suggest optimizing the MRF using a simulated annealing procedure [95], initialized using classic window-based DFD methods (Sec. 2.7.2).

**Defocus as diffusion.** Another approach to image restoration involves posing defocus in terms of a partial differential equation (PDE) for a diffusion process [39]. This strategy entails simulating the PDE on the more focused of the two images, until it becomes identical to the other image. Under this "simulation-based" inference, the time variable is related to the amount of relative blur. For isotropic diffusion, the formulation is equivalent to the simple isotropic heat equation, whereas for shift-variant diffusion, the anisotropy of the diffusion tensor characterizes the local variance of defocus.

**Deconvolution with occlusion.** Several recent DFD methods [42, 54, 77] have modeled occlusion effects in detail, following the richer reversed-projection model for defocus [16]. To make the reconstruction tractable, these methods assume a simplified scene model consisting of two smooth surfaces [42, 77], or use approximations for occluded defocus [54, 77].

Even though defocus under this occlusion model is no longer a simple convolutional operator, the simultaneous reconstruction of scene radiance, depth, and alpha mask is still amenable to image restoration techniques, using regularized gradient-based optimization [42, 54, 77].

**Information divergence.** All iterative deconvolution methods involve updating the estimated radiance and shape of the scene based on the discrepancy between the input images and the current synthetically defocused estimate. Based on several axioms and positivity constraints on scene radiance and the blur kernel, Favaro, *et al*. have argued that the only consistent measure of discrepancy is the information divergence, which generalizes the Kullback-Leibler (KL) divergence [38, 40, 62].

This discrepancy measure has been applied in the context of alternating minimization for surface and radiance [38, 40], as well as minimization by PDE gradient descent flow, using level set methods [62].

### 2.7.2 Depth from relative defocus

While direct deconvolution methods rely on simultaneously reconstructing the underlying scene radiance and depth, it is also possible to factor out the scene radiance, by considering the *relative* amount of defocus over a particular image window, between two different lens settings.

By itself, relative defocus is not enough to determine the depth of the scene, however the lens calibration may be used to resolve relative focus into absolute blur parameters, which can then be related to depth as before.

If one of the blur parameters is known in advance, the other blur parameter can be resolved by simple equation fitting. As Pentland describes, when one image is acquired with a pinhole aperture, *i.e*., $\sigma_1 = 0$, the relative blur directly determines the other blur parameter, *e.g*., according to Eq. (2.18) [91, 92].

In fact, the restriction that one of the images is a pinhole image can easily be relaxed, as the lens calibration provides an additional constraint between absolute blur parameters. For the thin lens model, Eq. (2.5) may be used to derive a linear constraint on the underlying blur parameters, *i.e*., $\sigma_1 = A\sigma_2 + B$, between any two lens settings [109]. Technically speaking, for

this linear constraint to be unique, the sign ambiguity in Eq. (2.5) must be resolved as described earlier in Sec. 2.7.

**Frequency-domain analysis.** The most common method of recovering relative defocus is by analyzing the relative frequency response in corresponding image windows. As shown below, the relative frequency response is invariant to the underlying scene radiance, and provides evidence about the level of relative defocus.

**Convolution ratio.** In the Fourier domain, the simple convolutional model of defocus given by Eq. (2.7) can be manipulated in an elegant way. Using the fact that convolution in the spatial domain corresponds to multiplication in the Fourier domain, we have

$$\begin{aligned}\frac{\mathcal{F}[\mathbf{I}_1]}{\mathcal{F}[\mathbf{I}_2]} &= \frac{\mathcal{F}[\mathbf{B}_{\sigma_1}] \cdot \mathcal{F}[\hat{\mathbf{I}}]}{\mathcal{F}[\mathbf{B}_{\sigma_2}] \cdot \mathcal{F}[\hat{\mathbf{I}}]} \\ &= \frac{\mathcal{F}[\mathbf{B}_{\sigma_1}]}{\mathcal{F}[\mathbf{B}_{\sigma_2}]} \ . \end{aligned} \tag{2.17}$$

This formula, also known as the *convolution ratio*, has the important feature of canceling all frequencies $\mathcal{F}[\hat{\mathbf{I}}]$ due to the underlying scene radiance [92, 129].

Thought of another way, the convolution ratio provides us with a relationship between the unknown defocus parameters, $\sigma_1$ and $\sigma_2$, that is invariant to the underlying scene. For example, by assuming a Gaussian defocus function as in Eqs. (2.13)–(2.14), the convolution ratio in Eq. (2.17) reduces to:

$$\sigma_2^2 - \sigma_1^2 = \frac{2}{\omega^2 + \nu^2} \ln\left(\frac{\mathcal{F}[\mathbf{I}_1](\omega, \nu)}{\mathcal{F}[\mathbf{I}_2](\omega, \nu)}\right) \ . \tag{2.18}$$

While other defocus functions may not admit such a simple closed-form solution, the convolution ratio will nevertheless describe a relationship between the blur parameters $\sigma_1$ and $\sigma_2$, *i.e.*, that can be expressed through a set of per-frequency lookup tables.

In theory, we can fully define the relative defocus, as in the prototypical Eq. (2.18), simply by considering the response of the defocused images at a single 2D frequency, $(\omega, \nu)$. However, using a fixed particular frequency can cause arbitrarily large errors and instability if the images do not contain sufficient energy in that frequency. Provided the defocus function is symmetric, for additional robustness we can integrate over radial frequency, $\lambda = \sqrt{\omega^2 + \nu^2}$, without affecting the relationship between relative blur and the convolution ratio [91, 92].

Note that an alternate version of the convolution ratio can be formulated using the total Fourier power, $\mathcal{P}(\omega, \nu) = |\mathcal{F}(\omega, \nu)|^2$, instead. This gives analogous equations for relative defo-

cus, but has the advantage that Parseval's theorem, $\iint |\mathbf{I}(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y = \frac{1}{4\pi^2} \iint |\mathcal{F}[\mathbf{I}](\omega,\nu)|^2 \,\mathrm{d}\omega\,\mathrm{d}\nu$, may be used to compute relative defocus more efficiently in the spatial domain [91, 109].

**Windowing effects.** By generalizing the convolution ratio to the windowed, locally fronto-parallel scene model described by Eq. (2.10), we obtain the more complicated formula,

$$\begin{aligned}\frac{\mathcal{F}[\mathbf{I}_1] * \mathcal{F}[\mathbf{W}_1]}{\mathcal{F}[\mathbf{I}_2] * \mathcal{F}[\mathbf{W}_2]} &= \frac{(\mathcal{F}[\mathbf{B}_{\sigma_1}] \cdot \mathcal{F}[\hat{\mathbf{I}}]) * \mathcal{F}[\mathbf{W}_1]}{(\mathcal{F}[\mathbf{B}_{\sigma_2}] \cdot \mathcal{F}[\hat{\mathbf{I}}]) * \mathcal{F}[\mathbf{W}_2]} \\ &\approx \frac{\mathcal{F}[\mathbf{B}_{\sigma_1}]}{\mathcal{F}[\mathbf{B}_{\sigma_2}]} \ . \end{aligned} \tag{2.19}$$

For the underlying scene radiance to cancel in this case, the windowing functions must be tightly band-limited in the Fourier domain, *i.e.*, $\mathcal{F}[\mathbf{W}_1] = \mathcal{F}[\mathbf{W}_2] \approx \delta$, which is only true for very large windows.

In addition to the previously described problems with windowing (Sec. 2.3.3), inverse filtering in Fourier domain presents additional difficulties due to finite-window effects [35]. Firstly, accurate spectral analysis requires large windows, which corresponds to low depth resolution or very smoothly varying scenes. Secondly, since windowing can be thought of as an additional convolution in the Fourier domain, it may cause zero-crossings in the Fourier domain to shift slightly, causing potentially large variations in convolution ratio. Finally, using same size windows in both images can lead to border artifacts, as the different levels of defocus imply that the actual source areas in $\hat{\mathbf{I}}$ are different.

**Tuned inverse filtering.** To mitigate the errors and instability caused by finite-width filters, one approach is to use filters specially tuned to the dominant frequencies in the image. Around the dominant frequencies, finite-width effects are negligible [129], however we do not know *a priori* which frequencies over an image window are dominant, or even if any exist.

A straightforward approach for identifying dominant frequencies, which provides an appealing formal invariance to surface radiance [43], is to use a large bank of tuned narrow-band filters, densely sampling the frequency domain [91, 128, 129]. Dominant frequencies can then be identified as filter responses of significant magnitude, and satisfying a stability criterion that detects contamination due to finite-width windowing artifacts [128, 129]. By assigning higher weights to the dominant frequencies, the depth estimates over all narrow-band filters may be aggregated, *e.g.*, using weighted least-squares regression.

Note that the uncertainty relation means that highly tuned filters with narrow response in

the frequency domain require large kernel support in the spatial domain. For a fixed window size in the spatial domain, Xiong and Shafer improve the resolution in the frequency domain by using additional moment-based filters, up to five times as many, to better model the spectrum within each narrow-band filter [129].

As Nayar, *et al*. suggest, another way to constrain the dominant frequencies in the scene is to use active illumination to project structured patterns onto the scene [81]. They propose using a checkerboard pattern, paired with a Laplacian-like focus operator that is tuned to the dominant frequency of the specific projected pattern.

**Broadband inverse filtering.** A contrasting approach to inverse filtering involves using broadband filters, that integrate over many different frequencies [109, 120]. Broadband filters lead to smaller windows in the spatial domain, and therefore to higher resolution; they are more computationally efficient, since less of them are required to estimate defocus; and they are more stable to low magnitude frequency responses. However, because defocus is not uniform over frequency (see Eq. (2.14), for example), the relative defocus estimated by integrating over a broad range of frequencies is potentially less accurate.

Watanabe and Nayar designed a set of three broadband filters for DFD, motivated as a higher-order expansion of the convolution ratio, Eq. (2.17), for a relatively small $7 \times 7$ spatial kernel [120]. In fact, they consider a normalized version of the convolution ratio instead,

$$\frac{\mathcal{F}[\mathbf{I}_1] - \mathcal{F}[\mathbf{I}_2]}{\mathcal{F}[\mathbf{I}_1] + \mathcal{F}[\mathbf{I}_2]} \approx \frac{\mathcal{F}[\mathbf{B}_{\sigma_1}] - \mathcal{F}[\mathbf{B}_{\sigma_2}]}{\mathcal{F}[\mathbf{B}_{\sigma_1}] + \mathcal{F}[\mathbf{B}_{\sigma_2}]} \quad . \tag{2.20}$$

constrained to $[-1, 1]$ for positive frequencies, and sharing the property that frequencies due to the underlying surface radiance cancel out [81, 120].

Watanabe and Nayar also suggest that it is important to pre-filter the image before inverse filtering, to remove any bias caused by the DC component, and to remove higher frequencies that violate the assumed monotonicity of the blur kernel [120]. Furthermore, to address the instability in low-texture regions, they define a confidence measure, derived using a perturbation analysis, and adaptively smooth the results until confidence meets some acceptable threshold throughout the image [120].

**Modeling relative defocus.** A different method for evaluating the relative defocus between two image windows is to explicitly model the operator representing relative defocus. Given such a model for relative defocus, we can potentially build a set of detectors corresponding to different

levels of relative defocus, producing output somewhat analogous to a focus measure (Sec. 2.5).

**Cubic polynomial patches.** Subbarao and Surya suggest modeling the ideal pinhole image as a cubic bivariate polynomial, that is, $\hat{\mathbf{I}} = \sum k_{m,n}\, x^m y^n$, with $m + n \leq 3$ [111]. Under this simple scene model, the convolution of Eq. (2.7) can be expressed in terms of low-order moments of the defocus function. Then by assuming the defocus function is radially symmetric, the deconvolution reduces to the analytic form, $\hat{\mathbf{I}} = \mathbf{I} - \frac{\sigma^2}{4} \cdot \nabla^2 \mathbf{I}$, which is equivalent to a well-known sharpening filter. Therefore the relative blur can be expressed analytically as

$$\mathbf{I}_2 - \mathbf{I}_1 = \frac{1}{4}(\sigma_2^2 - \sigma_1^2) \cdot \nabla^2 \left( \frac{\mathbf{I}_1 + \mathbf{I}_2}{2} \right) \quad , \tag{2.21}$$

for an arbitrary radially symmetric defocus function. Note that this expression contains no terms related to $\hat{\mathbf{I}}$, therefore it also provides invariance to scene radiance.

**Matrix-based deconvolution.** For the simple convolutional model of defocus described by Eq. (2.7), the convolution ratio (Sec. 2.7.2) can be easily reformulated in the spatial domain as

$$\mathbf{I}_2 = \mathbf{B}_\Delta * \mathbf{I}_1 \quad , \tag{2.22}$$

where $\mathbf{B}_\Delta$ is a kernel representing the relative defocus.

Because convolution is a linear operator, Eq. (2.22) can be expressed as matrix multiplication, where the matrix representing convolution is sparse, and has block-Toeplitz structure. While this suggests that $\mathbf{B}_\Delta$ may be recovered by matrix inversion, in the presence of noise, the problem is ill-posed and unstable. Ens and Lawrence propose performing this matrix inversion, but regularizing the solution by including a term that measures the fit of $\mathbf{B}_\Delta$ to a low-order polynomial [35].

Note that by manipulating the spatial domain convolution ratio, Eq. (2.22), we can obtain another expression for relative defocus, in terms of the underlying blur kernels,

$$\mathbf{B}_{\sigma_2} = \mathbf{B}_\Delta * \mathbf{B}_{\sigma_1} \quad . \tag{2.23}$$

Therefore, provided that the form of the blur kernel is known (Sec. 2.3), we can recover an explicit model of the relative defocus, $\mathbf{B}_\Delta$, by deconvolving Eq. (2.23). Ens and Lawrence suggest applying this method to recover $\mathbf{B}_\Delta$ over a range of different blur kernel pairs, corresponding to different depths [35]. Then each relative defocus hypothesis, $\mathbf{B}_\Delta$, can be evaluated by applying Eq. (2.22)

and measuring the least-squares error.

**Rank-based methods.** As discussed in Sec. 2.4, in one recent approach, the blur kernel may be recovered empirically as a set of rank-based approximations characterizing the relative defocus at particular calibrated depths [43]. This approach combines a large number of defocused images to recover a linear subspace describing an operator for relative defocus that provides invariance to the underlying scene radiance.

**Differential defocus.** Another way to model relative defocus is according to differential changes to the lens settings. As Subbarrao proposed in theory, the relative defocus can be fully specified by a differential analysis of the lens model [109].

Farid and Simoncelli realized an interesting version of differential DFD, by using specially designed pairs of optical filters that "directly" measure derivatives with respect to aperture size or viewpoint [36]. By comparing the image produced with one filter, and the spatial derivative of the image produced with another filter, they obtain a scale factor for every point, which can then be related to depth. This method relies on defocus, otherwise the scale factor will be undefined, and moreover it implicitly assumes a locally fronto-parallel scene over the extent of defocus [36].

### 2.7.3 Strong scene models

Assuming some prior knowledge, the scene may be reconstructed from as little as one defocused image. A strong model of the scene may also be used to simplify the depth estimation problem or to increase robustness.

**Sharp Reflectance Edges.** Various reconstruction methods from focus are based on the assumption that the in-focus scene contains perfectly sharp reflectance edges, *i.e.*, step discontinuities in surface albedo. Given this strong model of the scene, depth may be recovered by analyzing 1D intensity profiles across the blurred edges.

This approach was first proposed by Pentland for a single defocused image, as a qualitative measure over a sparse set of pixels containing detected edges [92]. The analysis was subsequently generalized for rotationally symmetric blur kernels [110], where it was shown that the radius of gyration (Sec. 2.5.1) is simply $\sqrt{2}$ times the second moment of the line spread function for a sharp edge.

Asada, *et al.* proposed a more robust method for detecting sharp edges and their depths, using a dense number of focus settings [15]. Under the assumption of a rotationally symmet-

ric blur model, constant-intensity lines may be fit in the vicinity of each edge, where depth is determined by the intersection of these lines.

**Confocal lighting.** As already noted, active illumination can be used constrain the frequency characteristics of the scene, and increase the robustness of estimating the relative defocus (Sec. 2.7.2). However, an even stronger use of active illumination is confocal lighting, which involves selectively illuminating the equifocal surface with a light source sharing the same optical path as the lens [121].

By using a large aperture together with confocal lighting, parts of the scene away from the equifocal surface are both blurred and dark, which greatly enhances contrast [73, 121]. In theory, we can directly obtain cross-sections of the scene from such images at different focus settings, and assemble them into a 3D volumetric model.

Levoy, *et al*. suggest a version of confocal imaging implemented on macro-scale with opaque objects [73]. In their design, a single camera and projector share the same optical path using a beam splitter, and they use an array of mirrors to create 16 virtual viewpoints. Then, multi-pixel tiles at a given focal depth are illuminated according to coded masks, by focusing the projector from the virtual viewpoints. Although they obtain good results with their system for matting, the depth of field is currently too large to provide adequate resolution for 3D reconstruction.

### 2.7.4 Analysis

**Feasibility of DFD.** Favaro, *et al*. provide a basic result that if the radiance of the scene can be arbitrarily controlled, for example, by active illumination, then any piecewise-smooth surface can be distinguished from all others, *i.e*., fully reconstructed, from a set of defocused images [38].

With the "complexity" of scene radiance formalized in terms of the degree of a 2D linear basis, it can be shown that two piecewise-smooth surfaces can theoretically be distinguished up to a resolution that depends on this complexity, with further limitations due to the optics [38].

**Optimal interval for DFD.** Using a perturbation analysis of thin lens model, assuming a pillbox defocus function, Schechner and Kiryati showed that the optimal interval between the two focus settings, with respect to perturbations at the Nyquist frequency, corresponds to the depth of field [99]. For smaller intervals no frequency will satisfy optimality, whereas for larger intervals the Nyquist frequency will be suboptimal, but some lower frequency will be optimal with respect to perturbations.

Shih, *et al*. give an alternate analysis of optimality to perturbations in information-theoretic terms, assuming a symmetric Gaussian defocus function [102]. According to their analysis, the lowest-variance unbiased estimator from a pair of defocused images is attained when the corresponding levels of blur are related by $\sigma_1 = \sqrt{3/2}\,\sigma_2$. However, this result is difficult to apply in practice, since knowledge of these blur parameters implies that depth has already been recovered.

## 2.8 Compositing and Resynthesis

Although most methods for processing defocused images have concentrated on 3D reconstruction, others have explored synthesizing new images from this input as well [10, 29, 51, 54, 87]. A growing interest in this application has also motivated specialized imaging designs that capture representations enabling the photographer to refocus or manipulate other camera parameters after the capture session [61, 69, 85, 115].

**Image fusion for extended depth of field.** The most basic image synthesis application for multiple defocused images is to synthetically create a composite image, in which the whole scene is in-focus [10, 51, 87, 97, 108, 130, 135]. This application is of special interest to macro-scale photographers, because in this domain the depth of field is so limited that capturing multiple photos with different settings are often required just to span the desired depth range of the subject.

Classic methods of this type involve applying a blind focus measure (Sec. 2.5.2) to a set of differently focused images, followed by a hard, winner-take-all decision rule over various image scales, selecting the pixels determined to be the most in-focus [87, 135]. More recently, this approach has been extended to incorporate adaptively-sized and variously oriented windows for the computation of the defocus measure [97, 108]. Such methods typically exhibit strong artifacts at depth discontinuities, and are biased toward noisy composites, since the high frequencies associated with noise are easily mistaken for in-focus texture.

More successful recent methods for image fusion are based on first performing depth-from-focus (Sec. 2.6) in an MRF framework [10, 130]. An advantage of this approach is that by favoring global piecewise smoothness, over-fitting to image noise can be avoided. Another important feature for generating a visually realistic composite is the use of gradient-based blending [10], which greatly reduces compositing artifacts at depth discontinuities.

Note that while other 3D reconstruction methods based on image restoration from defocus (Sec. 2.7.1) implicitly recover an underlying in-focus representation of the scene as well, these

methods are not designed with its display as a specific goal.

**Resynthesis with new camera parameters.** In a recent application of the convolution ratio (Sec. 2.7.2), Chaudhuri suggested a nonlinear interpolation method for "morphing" between two defocused images taken with different settings [29]. Although the formulation is elegant, the method shares the limitations of other additive window-based methods (Sec. 2.3.3) and does not address the inherent depth ambiguity described by Eq. (2.5). In particular, because the interpolation does not allow the possibility of the in-focus setting lying between the focus settings of the input images, the synthesized results may be physically inconsistent.

Another resynthesis application for defocused images is to synthetically increase the level of defocus, to reproduce the shallow depth of field found in large-aperture SLR photos. As Bae and Durand show, for the purpose of this simple application, defocus can be estimated sufficiently well just from cues in a single image [18].

In general, any depth-from-defocus method yielding both a depth map and the underlying in-focus radiance (Sec. 2.7.1) can be exploited to resynthesize new images with simulated camera settings (*e.g.*, refocusing), according to the assumed forward image formation model. One of the earliest methods to consider this problem involved implicitly decomposing the scene into additive transparent layers [14, 65]. A more recent approach used a layer-based scene model as well, but incorporates a detailed model of occlusion at depth discontinuities [54].

# Chapter 3

# Confocal Stereo

There is nothing worse than a sharp image of a fuzzy concept.
Ansel Adams (1902–1984)

In this chapter, we present *confocal stereo*, a new method for computing 3D shape by controlling the focus and aperture of a lens. The method is specifically designed for reconstructing scenes with high geometric complexity or fine-scale texture. To achieve this, we introduce the *confocal constancy* property, which states that as the lens aperture varies, the pixel intensity of a visible in-focus scene point will vary in a scene-*independent* way, that can be predicted by prior radiometric lens calibration. The only requirement is that incoming radiance within the cone subtended by the largest aperture is nearly constant. First, we develop a detailed lens model that factors out the distortions in high resolution SLR cameras (12 MP or more) with large-aperture lenses (*e.g.*, f/1.2). This allows us to assemble an $A \times F$ aperture-focus image (AFI) for each pixel, that collects the undistorted measurements over all $A$ apertures and $F$ focus settings. In the AFI representation, confocal constancy reduces to color comparisons within regions of the AFI, and leads to focus metrics that can be evaluated separately for each pixel. We propose two such metrics and present initial reconstruction results for complex scenes, as well as for a scene with known ground-truth shape.

## 3.1 Introduction

Recent years have seen many advances in the problem of reconstructing complex 3D scenes from multiple photographs [45, 57, 137]. Despite this progress, however, there are many common scenes for which obtaining detailed 3D models is beyond the state of the art. One such

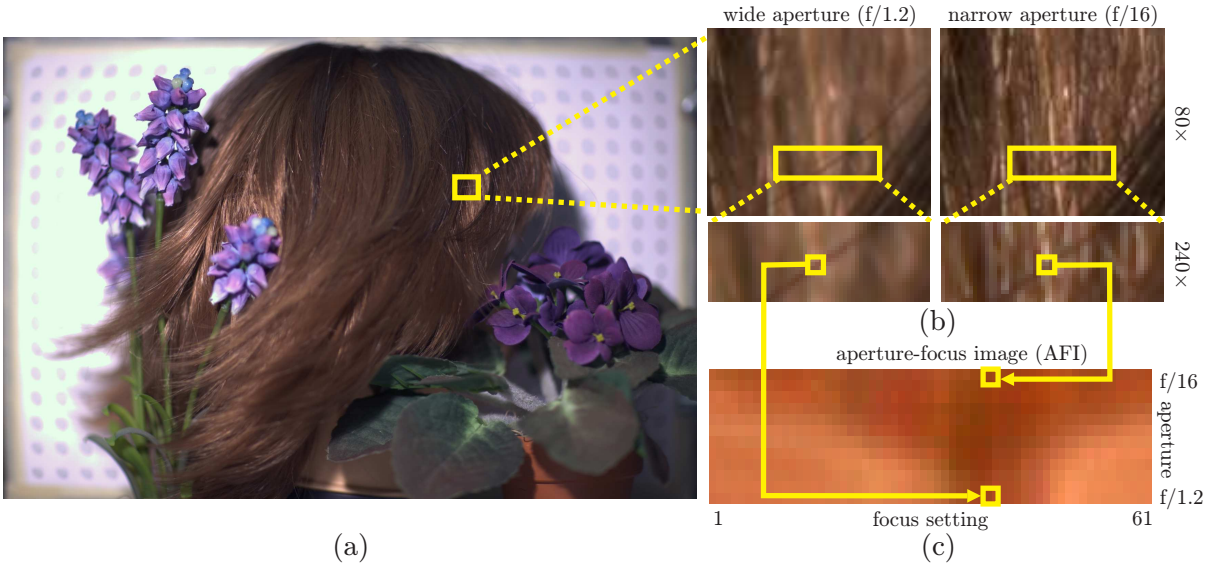



**Figure 3.1:** (a) Wide-aperture image of a complex scene. (b) *Left:* Successive close-ups of a region in (a), showing a single in-focus strand of hair. *Right:* Narrow-aperture image of the same region, with everything in focus. Confocal constancy tells us that the intensity of in-focus pixels (*e.g.*, on the strand) changes predictably between these two views. (c) The aperture-focus image (AFI) of a pixel near the middle of the strand. A column of the AFI collects the intensities of that pixel as the aperture varies with focus fixed.

class includes scenes that contain very high levels of geometric detail, such as hair, fur, feathers, miniature flowers, *etc*. These scenes are difficult to reconstruct for a number of reasons—they create complex 3D arrangements not directly representable as a single surface; their images contain fine detail beyond the resolution of common video cameras; and they create complex self-occlusion relationships. As a result, many approaches either side-step the reconstruction problem [45], require a strong prior model for the scene [89, 122], or rely on techniques that approximate shape at a coarse level.

Despite these difficulties, the high-resolution sensors in today's digital cameras open the possibility of imaging complex scenes at a very high level of detail. With resolutions surpassing 12Mpixels, even individual strands of hair may be one or more pixels wide (Fig. 3.1a,b). In this chapter, we explore the possibility of reconstructing static scenes of this type using a new method called *confocal stereo*, which aims to compute depth maps at sensor resolution. Although the method applies equally well to low-resolution settings, it is designed to exploit the capabilities of high-end digital SLR cameras and requires no special equipment besides the camera and a laptop. The only key requirement is the ability to actively control the aperture, focus setting, and exposure time of the lens.

At the heart of our approach is a property we call *confocal constancy*, which states that as

the lens aperture varies, the pixel intensity of a visible in-focus scene point will vary in a scene-*independent* way, that can be predicted by prior radiometric lens calibration. To exploit confocal constancy for reconstruction, we develop a detailed lens model that factors out the geometric and radiometric distortions observable in high resolution SLR cameras with large-aperture lenses (*e.g.*, f/1.2). This allows us to assemble an $A \times F$ *aperture-focus image (AFI)* for each pixel, that collects the undistorted measurements over all $A$ apertures and $F$ focus settings (Fig 3.1c). In the AFI representation, confocal constancy reduces to color comparisons within regions of the AFI and leads to focus metrics that can be evaluated separately for each pixel.

Our work has four main contributions. First, unlike existing depth from focus or depth from defocus methods, our confocal constancy formulation shows that we can assess focus without modeling a pixel's spatial neighborhood or the blurring properties of a lens. Second, we show that depth from focus computations can be reduced to pixelwise intensity comparisons, in the spirit of traditional stereo techniques. Third, we introduce the aperture-focus-image representation as a basic tool for focus- and defocus-based 3D reconstruction. Fourth, we show that together, confocal constancy and accurate image alignment lead to a reconstruction algorithm that can compute depth maps at resolutions not attainable with existing techniques. To achieve all this, we also develop a method for the precise geometric and radiometric alignment of high-resolution images taken at multiple focus and aperture settings, that is particularly suited for professional-quality cameras and lenses, where the standard thin-lens model breaks down.

We begin this chapter by discussing the relation of this work to current approaches for reconstructing scenes that exploit defocus in wide-aperture images. Sec. 3.3 describes our generic imaging model and introduces the property of confocal constancy. Sec. 3.4 gives a brief overview of how we exploit this property for reconstruction and Secs. 3.5–3.6 discuss the radiometric and geometric calibration required to relate high resolution images taken with different lens settings. In Sec. 3.7 we show how the AFI for each pixel can be analyzed independently to estimate depth, using both confocal constancy and its generalization. Finally, Sec. 3.8 presents experimental results using images of complex real scenes, and one scene for which ground truth has been recovered.

## 3.2 Related Work

Our method builds on five lines of recent work—depth from focus, depth from defocus, shape from active illumination, camera calibration, and synthetic aperture imaging. We briefly discuss their relation to this work below.

**Depth from focus.** Our approach can be thought of as a depth from focus method, in that we assign depth to each pixel by selecting the focus setting that maximizes a focus metric for that pixel's AFI. Classic depth from focus methods collect images at multiple focus settings and define metrics that measure sharpness over a small spatial window surrounding the pixel [30, 64, 80]. This implicitly assumes that depth is approximately constant for all pixels in that window. In contrast, our criterion depends on measurements at a single pixel and requires manipulating a second, independent camera parameter (*i.e.*, aperture). As a result, we can recover much sharper geometric detail than window-based methods, and also recover depth with more accuracy near depth discontinuities. The tradeoff is that our method requires us to capture more images than other depth from focus methods.

**Depth from defocus.** Many depth from defocus methods directly evaluate defocus over spatial windows, *e.g.*, by fitting a convolutional model of defocus to images captured at different lens settings [43, 49, 92, 111, 120, 129]. Spatial windowing is also implicit in recent depth from defocus methods based on deconvolving a single image, with the help of coded apertures and natural image statistics [69, 115]. As a result, none of these methods can handle scenes with dense discontinuities like the ones we consider. Moreover, while depth from defocus methods generally exploit basic models of defocus, the models used do not capture the complex blurring properties of multi-element, wide-aperture lenses, which can adversely affect depth computations.

Although depth from defocus methods have taken advantage of the ability to control camera aperture, this has generally been used as a substitute for focus control, so the analysis remains essentially the same [49, 92, 111]. An alternative form of aperture control involves using specially designed pairs of optical filters in order to compute derivatives with respect to aperture size or viewpoint [36], illuminating the connection between defocus-based methods and small-baseline stereo [36, 99]. Our method, on the other hand, is specifically designed to exploit image variations caused by changing the aperture in the standard way.

A second class of depth from defocus methods formulates depth recovery as an iterative global energy minimization problem, simultaneously estimating depth and in-focus radiance at all pixels [22, 38, 39, 42, 54, 62, 77, 95]. Some of the recent methods in this framework model defocus in greater detail to better handle occlusion boundaries [22, 42, 54, 77] (see Chapter 4), but rely on the occlusion boundaries being smooth. Unfortunately, these minimization-based methods are prone to many local minima, their convergence properties are not completely understood, and they rely on smoothness priors that limit the spatial resolution of recovered depth maps.

Compared to depth from defocus methods, which may require as little as a single image [69, 115], our method requires us to capture many more images. Again, the tradeoff is that our method provides us with the ability to recover pixel-level depth for fine geometric structures, which would not otherwise be possible.

**Shape from active illumination.** Since it does not involve actively illuminating the scene, our reconstruction approach is a "passive" method. Several methods use active illumination (*i.e.*, projectors) to aid defocus computations. For example, by projecting structured patterns onto the scene, it is possible to control the frequency characteristics of defocused images, reducing the influence of scene texture [38, 79, 81]. Similarly, by focusing the camera and the projected illumination onto the same scene plane, confocal microscopy methods are able to image (and therefore reconstruct) transparent scenes one slice at a time [121]. This approach has also been explored for larger-scale opaque scenes [73].

Most recently, Zhang and Nayar developed an active illumination method that also computes depth maps at sensor resolution [132]. To do this, they evaluate the defocus of patterns projected onto the scene using a metric that also relies on single-pixel measurements. Their approach can be thought of as orthogonal to our own, since it projects multiple defocused patterns instead of controlling aperture. While their preliminary work has not demonstrated the ability to handle scenes of the spatial complexity discussed here, it may be possible to combine aperture control and active illumination for more accurate results. In practice, active illumination is most suitable for darker environments, where the projector is significantly brighter than the ambient lighting.

**Geometric and radiometric lens calibration.** Because of the high image resolutions we employ (12Mpixels or more) and the need for pixel-level alignment between images taken at multiple lens settings, we model detailed effects that previous methods were not designed to handle. For example, previous methods account for radiometric variation by normalizing spatial image windows by their mean intensity [92, 111], or by fitting a global parametric model such as a cosine-fourth falloff [63]. To account for subtle radiometric variations that occur in multi-element, off-the-shelf lenses, we use a data-driven, non-parametric model that accounts for the camera response function [31, 50] as well as slight temporal variations in ambient lighting. Furthermore, most methods for modeling geometric lens distortions due to changing focus or zoom setting rely on simple magnification [15, 30, 81, 119] or radial distortion models [124], which are not sufficient to achieve sub-pixel alignment of high resolution images.

**Synthetic aperture imaging.** While real lenses integrate light over wide apertures in a continuous fashion, multi-camera systems can be thought of as a discretely-sampled synthetic aperture that integrates rays from the light field [74]. Various such systems have been proposed in recent years, including camera arrays [61, 74], virtual camera arrays simulated using mirrors [73], and arrays of lenslets in front of a standard imaging sensor [9, 85]. Our work can be thought of as complementary to these methods since it does not depend on having a single physical aperture; in principle, it can be applied to synthetic apertures as well.

## 3.3 Confocal Constancy

Consider a camera whose lens contains multiple elements and has a range of known focus and aperture settings. We assume that no information is available about the internal components of this lens (*e.g.*, the number, geometry, and spacing of its elements). We therefore model the lens as a "black box" that redirects incoming light toward a fixed sensor plane and has the following idealized properties:

- **Negligible absorption:** light that enters the lens in a given direction is either blocked from exiting or is transmitted with no absorption.
- **Perfect focus:** for every 3D point in front of the lens there is a unique focus setting that causes rays through the point to converge to a single pixel on the sensor plane.
- **Aperture-focus independence:** the aperture setting controls only which rays are blocked from entering the lens; it does not affect the way that light is redirected.

These properties are well approximated by lenses used in professional photography applications[1]. Here we use such a lens to collect images of a 3D scene for $A$ aperture settings, $\{\alpha_1, \ldots, \alpha_A\}$, and $F$ focal settings, $\{f_1, \ldots, f_F\}$. This acquisition produces a 4D set of pixel data, $\mathbf{I}_{\alpha f}(x, y)$, where $\mathbf{I}_{\alpha f}$ is the image captured with aperture $\alpha$ and focal setting $f$. As in previous defocus-based methods, we assume that the camera and scene are stationary during the acquisition [64, 92, 132].

Suppose that a 3D point $\mathbf{p}$ on an opaque surface is in perfect focus in image $\mathbf{I}_{\alpha f}$ and suppose that it projects to pixel $(x, y)$. In this case, the light reaching the pixel is restricted to a cone from $\mathbf{p}$ that is determined by the aperture setting (Fig. 3.2). For a sensor with a linear response, the intensity $\mathbf{I}_{\alpha f}(x, y)$ measured at the pixel is proportional to the irradiance, namely the integral

[1] There is a limit, however, on how close points can be and still be brought into focus for real lenses, restricting the 3D workspace that can be reconstructed.

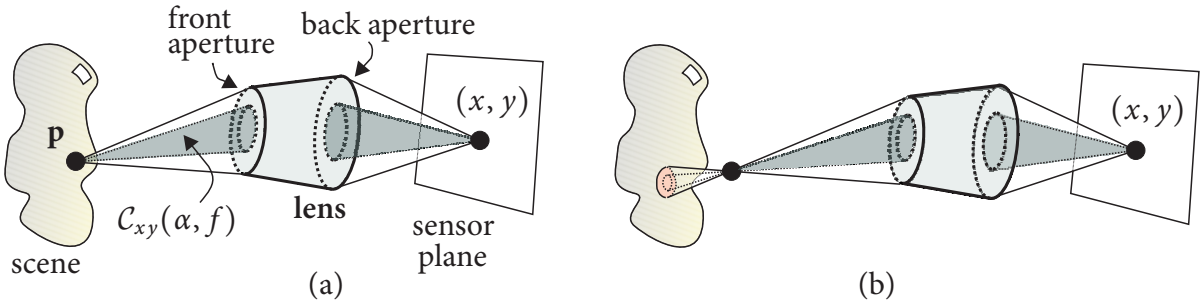



**Figure 3.2:** Generic lens model. (a) At the perfect focus setting of pixel $(x, y)$, the lens collects outgoing radiance from a scene point **p** and directs it toward the pixel. The 3D position of point **p** is uniquely determined by pixel $(x, y)$ and its perfect focus setting. The shaded cone of rays, $\mathcal{C}_{xy}(\alpha, f)$, determines the radiance reaching the pixel. This cone is a subset of the cone subtended by **p** and the front aperture because some rays may be blocked by internal components of the lens, or by its back aperture. (b) For out-of-focus settings, the lens integrates outgoing radiance from a region of the scene.

of outgoing radiance over the cone,

$$\mathbf{I}_{\alpha f}(x, y) \;=\; \kappa \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} L(\mathbf{p}, \omega)\, \mathrm{d}\omega \;\;, \tag{3.1}$$

where $\omega$ measures solid angle, $L(\mathbf{p}, \omega)$ is the radiance for rays passing through **p**, $\kappa$ is a constant that depends only on the sensor's response function [31, 50], and $\mathcal{C}_{xy}(\alpha, f)$ is the cone of rays that reach $(x, y)$. In practice, the apertures on a real lens correspond to a nested sequence of cones, $\mathcal{C}_{xy}(\alpha_1, f) \subset \ldots \subset \mathcal{C}_{xy}(\alpha_A, f)$, leading to a monotonically-increasing intensity at the pixel (given equal exposure times).

If the outgoing radiance at the in-focus point **p** remains constant within the cone of the largest aperture, *i.e.*, $L(\mathbf{p}, \omega) = L(\mathbf{p})$, and if this cone does not intersect the scene elsewhere, the relation between intensity and aperture becomes especially simple. In particular, the integral of Eq. (3.1) disappears and the intensity for aperture $\alpha$ is proportional to the solid angle subtended by the associated cone, *i.e.*,

$$\mathbf{I}_{\alpha f}(x, y) \;=\; \kappa \,\|\, \mathcal{C}_{xy}(\alpha, f) \,\|\, L(\mathbf{p}) \;\;, \tag{3.2}$$

where $\|\, \mathcal{C}_{xy}(\alpha, f) \,\| = \int_{\mathcal{C}_{xy}(\alpha, f)} \mathrm{d}\omega$. As a result, the ratio of intensities at an in-focus point for two different apertures is a scene-independent quantity:

**Confocal Constancy Property**

$$\frac{\mathbf{I}_{\alpha f}(x, y)}{\mathbf{I}_{\alpha_1 f}(x, y)} \;=\; \frac{\|\, \mathcal{C}_{xy}(\alpha, f) \,\|}{\|\, \mathcal{C}_{xy}(\alpha_1, f) \,\|} \;\stackrel{\text{def}}{=}\; \mathbf{R}_{xy}(\alpha, f) \;\;. \tag{3.3}$$

Intuitively, the constant of proportionality, $\mathbf{R}_{xy}(\alpha, f)$, describes the relative amount of light received from an in-focus scene point for a given aperture. This constant, which we call the *relative exitance* of the lens, depends on lens internal design (front and back apertures, internal elements, *etc.*) and varies in general with aperture, focus setting, and pixel position on the sensor plane. Thus, relative exitance incorporates vignetting and other similar radiometric effects that do not depend on the scene.

Confocal constancy is an important property for evaluating focus for four reasons. First, it holds for a very general lens model that covers the complex lenses commonly used with high-quality SLR cameras. Second, it requires no assumptions about the appearance of out-of-focus points. Third, it holds for scenes with general reflectance properties, provided that radiance is nearly constant over the cone subtended by the largest aperture.[2] Fourth, and most important, it can be evaluated at *pixel resolution* because it imposes no requirements on the spatial layout (*i.e.*, depths) of points in the neighborhood of **p**.

## 3.4 The Confocal Stereo Procedure

Confocal constancy allows us to decide whether or not the point projecting to a pixel $(x, y)$ is in focus by comparing the intensities $\mathbf{I}_{\alpha f}(x, y)$ for different values of aperture $\alpha$ and focus $f$. This leads to the following reconstruction procedure (Fig. 3.3):

1. **(Relative exitance estimation)** Compute the relative exitance of the lens for the $A$ apertures and $F$ focus settings (Sec. 3.5).

2. **(Image acquisition)** For each of the $F$ focus settings, capture an image of the scene for each of the $A$ apertures.

3. **(Image alignment)** Warp the captured images to ensure that a scene point projects to the same pixel in all images (Sec. 3.6).

4. **(AFI construction)** Build an $A \times F$ aperture-focus image for each pixel, that collects the pixel's measurements across all apertures and focus settings.

5. **(Confocal constancy evaluation)** For each pixel, process its AFI to find the focus setting that best satisfies the confocal constancy property (Sec. 3.7).

[2] For example, an aperture with an effective diameter of 70 mm located 1.2 m from the scene corresponds to 0.5 % of the hemisphere, or a cone whose rays are less than 3.4° apart.

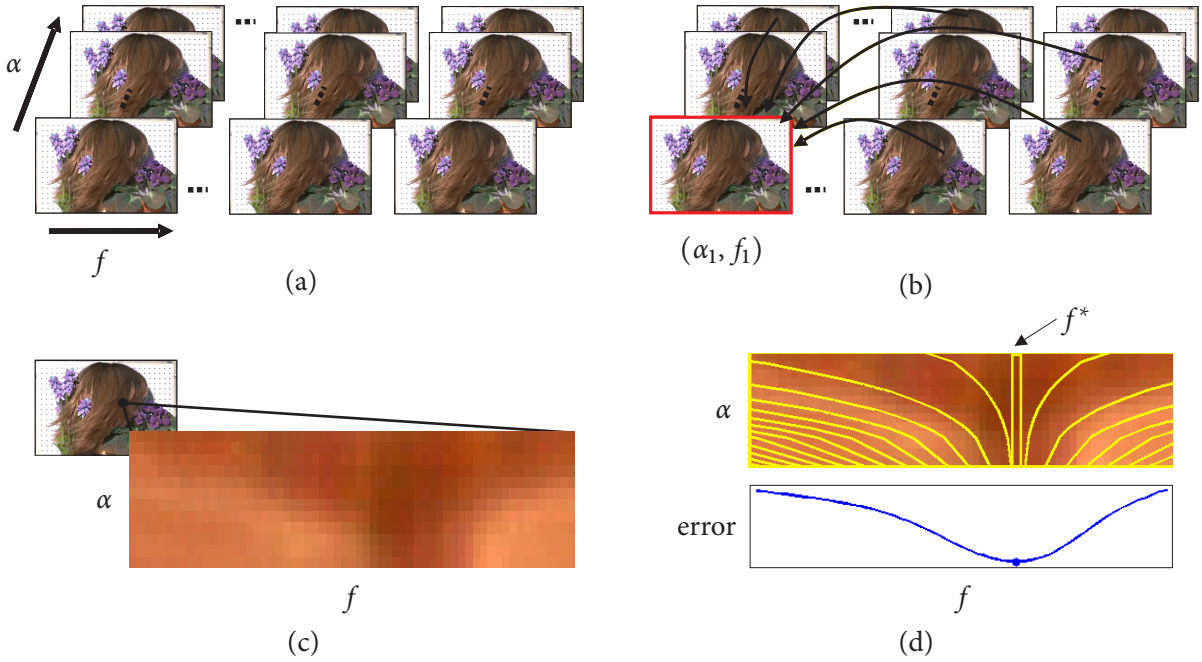


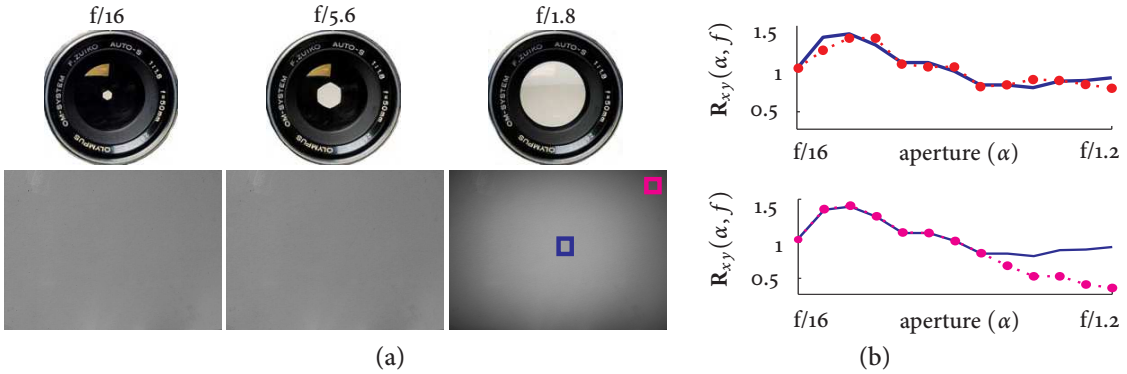

**Figure 3.3:** Overview of confocal stereo: (a) Acquire $A \times F$ images over $A$ apertures and $F$ focus settings. (b) Align all images to the reference image, taking into account both radiometric calibration (Sec. 3.5) and geometric distortion (Sec. 3.6). (c) Build the $A \times F$ aperture-focus image (AFI) for each pixel. (d) Process the AFI to find the best in-focus setting (Sec. 3.7).

## 3.5 Relative Exitance Estimation

In order to use confocal constancy for reconstruction, we must be able to predict how changing the lens aperture affects the appearance of scene points that are in focus. Our approach is motivated by three basic observations. First, the apertures on real lenses are non-circular and the f-stop values describing them only approximate their true area (Fig. 3.4a,b). Second, when the effective aperture diameter is a relatively large fraction of the camera-to-object distance, the solid angles subtended by different 3D points in the workspace can differ significantly.[3] Third, vignetting and off-axis illumination effects cause additional variations in the light gathered from different in-focus points [63, 105] (Fig. 3.4b).

To deal with these issues, we explicitly compute the relative exitance of the lens, $\mathbf{R}_{xy}(\alpha, f)$, for all apertures $\alpha$ and for a sparse set of focal settings $f$. This can be thought of as a scene-independent radiometric lens calibration step that must be performed just once for each lens. In practice, this allows us to predict aperture-induced intensity changes to within the sensor's noise level (*i.e.*, within 1–2 gray levels), and enables us to analyze potentially small intensity

[3] For a 70 mm diameter aperture, the solid angle subtended by scene points 1.1–1.2 m away can vary up to 10%.




(a) (b)

**Figure 3.4:** (a) Images of an SLR lens showing variation in aperture shape with corresponding images of a diffuse plane. (b) *Top:* comparison of relative exitances for the central pixel indicated in (a), as measured using Eq. (3.3) (solid graph), and as approximated using the f-stop values (dotted) according to $\mathbf{R}_{xy}(\alpha, f) = \alpha_1^2/\alpha^2$ [31]. *Bottom:* comparison of the central pixel (solid) with the corner pixel (dotted) indicated in (a). The agreement is good for narrow apertures (*i.e.*, high f-stop values), but for wider apertures, spatially-varying effects are significant.

variations due to focus. For quantitative validation of our radiometric calibration method, see Appendix A.

To compute relative exitance for a focus setting $f$, we place a diffuse white plane at the in-focus position and capture one image for each aperture, $\alpha_1, \ldots, \alpha_A$. We then apply Eq. (3.3) to the luminance values of each pixel $(x, y)$ to recover $\mathbf{R}_{xy}(\alpha_i, f)$. To obtain $\mathbf{R}_{xy}(\alpha_i, f)$ for focus settings that span the entire workspace, we repeat the process for multiple values of $f$ and use interpolation to compute the in-between values. Since Eq. (3.3) assumes that pixel intensity is a linear function of radiance, we transform all images according to the inverse of the sensor response function, which we recover using standard techniques from the high dynamic range literature [31, 50].

Note that in practice, we manipulate the exposure time in conjunction with the aperture setting $\alpha$, to keep the total amount of light collected roughly constant and prevent unnecessary pixel saturation. Exposure time can be modeled as an additional multiplicative factor in the image formation model, Eq. (3.1), and does not affect the focusing behavior of the lens.[4] Thus, we can fold variation in exposure time into the calculation of $\mathbf{R}_{xy}(\alpha_i, f)$, provided that we vary the exposure time in the same way for both the calibration and test sequences.

**Global lighting correction.** While the relative exitance need only be computed once for a given lens, we have observed that variations in ambient lighting intensity over short time in-

[4] A side-effect of manipulating the exposure time is that noise characteristics will change with varying intensity [56], however this phenomenon does not appear to be significant in our experiments.

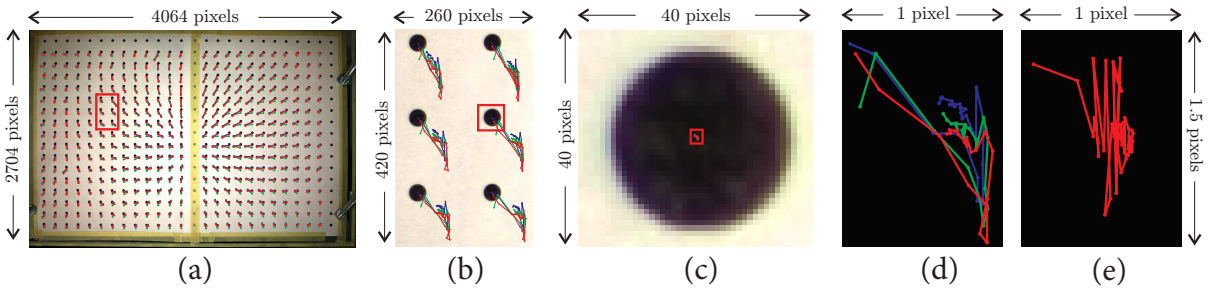



**Figure 3.5:** (a–e) To evaluate stochastic lens distortions, we computed centroids of dot features for images of a static calibration pattern. (a–d) Successive close-ups of a centroid's trajectory for three cycles (red, green, blue) of the 23 aperture settings. In (a–b) the trajectories are magnified by a factor of 100. As shown in (d), the trajectory, while stochastic, correlates with aperture setting. (e) Trajectory for the centroid of (c) over 50 images with the same lens settings.

tervals can be significant (especially for fluorescent tubes, due to voltage fluctuations). This prevents directly applying the relative exitance computed during calibration to a different sequence.

To account for this effect, we model lighting variation as an unknown multiplicative factor that is applied globally to each captured image. To factor out lighting changes, we renormalize the images so that the total intensity of a small patch at the image center remains constant over the image sequence. In practice, we use a patch that is a small fraction of the image (roughly 0.5 % of the image area), so that aperture-dependent effects such as vignetting can be ignored, and we take into account only pixels that are unsaturated for every lens setting.

## 3.6 High-Resolution Image Alignment

The intensity comparisons needed to evaluate confocal constancy are only possible if we can locate the projection of the same 3D point in multiple images taken with different settings. The main difficulty is that real lenses map in-focus 3D points onto the image plane in a non-linear fashion that cannot be predicted by ordinary perspective projection. To enable cross-image comparisons, we develop an alignment procedure that reverses these non-linearities and warps the input images to make them consistent with a reference image (Fig. 3.3b).

Since our emphasis is on reconstructing scenes at the maximum possible spatial resolution, we aim to model real lenses with enough precision to ensure sub-pixel alignment accuracy. This task is especially challenging because at resolutions of 12 MP or more, we begin to approach the optical and mechanical limits of the camera. In this domain, the commonly-used thin lens (*i.e.*, magnification) model [16, 30, 39, 41, 42, 81] is insufficient to account for observed distortions.

### 3.6.1 Deterministic second-order radial distortion model

To model geometric distortions caused by the lens optics, we use a model with $F+5$ parameters for a lens with $F$ focal settings. The model expresses deviations from an image with reference focus setting $f_1$ as an additive image warp consisting of two terms—a pure magnification term $m_f$ that is specific to focus setting $f$, and a quadratic distortion term that amplifies the magnification:

$$\mathbf{w}_f^{\mathrm{D}}(x,y) = \left[\, m_f + m_f(f-f_1)(k_0 + k_1 r + k_2 r^2) - 1 \,\right] \cdot \left[ (x,y) - (x_c, y_c) \right] \;, \tag{3.4}$$

where $k_0, k_1, k_2$ are the quadratic distortion parameters, $(x_c, y_c)$ is the estimated image center, and $r = \|(x,y) - (x_c, y_c)\|$ is the radial displacement.[5] Note that when the quadratic distortion parameters are zero, the model reduces to pure magnification, as in the thin lens model.

It is a standard procedure in many methods [67, 124] to model radial distortion using a polynomial of the radial displacement, $r$. A difference in our model is that the quadratic distortion term in Eq. (3.4) incorporates a linear dependence on the focus setting as well, consistent with more detailed calibration methods involving distortion components related to distance [46]. In our empirical tests, we have found that this term is necessary to obtain sub-pixel registration at high resolutions.

### 3.6.2 Stochastic first-order distortion model

We were surprised to find that significant misalignments can occur even when the camera is controlled remotely without any change in settings and is mounted securely on an optical table (Fig. 3.5e). While these motions are clearly stochastic, we also observed a reproducible, aperture-dependent misalignment of about the same magnitude (Fig. 3.5a–d), which corresponded to slight but noticeable changes in viewpoint. In order to achieve sub-pixel alignment, we approximate these motions by a global 2D translation, estimated independently for every image:

$$\mathbf{w}_{\alpha f}^{\mathrm{S}}(x,y) = \mathbf{t}_{\alpha f} \;. \tag{3.5}$$

We observed these motions with two different Canon lenses and three Canon SLR cameras, with no significant difference using mirror-lockup mode. We hypothesize that this effect is caused by additive random motion due to camera vibrations, plus variations in aperture shape and its

[5] Since our geometric distortion model is radial, the estimated image center has zero displacement over focus setting, *i.e.*, $\mathbf{w}_f^{\mathrm{D}}(x_c, y_c) = (0,0)$ for all $f$.

center point.

Note that while the geometric image distortions have a stochastic component, the correspondence itself is deterministic: given two images taken at two distinct camera settings there is a unique correspondence between their pixels.

### 3.6.3 Offline geometric lens calibration

We recover the complete distortion model of Eqs. (3.4)–(3.5) in a single optimization step, using images of a calibration pattern taken over all $F$ focus settings at the narrowest aperture, $\alpha_1$. This optimization simultaneously estimates the $F + 5$ parameters of the deterministic model and the $2F$ parameters of the stochastic model. To do this, we solve a non-linear least squares problem that minimizes the squared reprojection error over a set of features detected on the calibration pattern:

$$E(x_c, y_c, \mathbf{m}, \mathbf{k}, \mathbf{T}) = \sum_{(x,y)} \sum_f \left\| \mathbf{w}_f^{\mathrm{D}}(x, y) + \mathbf{w}_{\alpha_1 f}^{\mathrm{S}}(x, y) - \Delta_{\alpha_1 f}(x, y) \right\|^2 , \tag{3.6}$$

where $\mathbf{m}$ and $\mathbf{k}$ are the vectors of magnification and quadratic parameters, respectively; $\mathbf{T}$ collects stochastic translations; and $\Delta_{\alpha_1 f}(x, y)$ is the displacement between a feature location at focus setting $f$ and its location at the reference focus setting, $f_1$.

To avoid being trapped in a local minimum, we initialize the optimization with suitable estimates for $(x_c, y_c)$ and $\mathbf{m}$, and initialize the other distortion parameters to zero. To estimate the image center $(x_c, y_c)$, we fit lines through each feature track across focus setting, and then compute their "intersection" as the point minimizing the sum of distances to these lines. To estimate the magnifications $\mathbf{m}$, we use the regression suggested by Willson and Shafer [127] to aggregate the relative expansions observed between pairs of features.

In practice, we use a planar calibration pattern consisting of a grid of about $25 \times 15$ circular black dots on a white background (Fig. 3.5). We roughly localize the dots using simple image processing and then compute their centroids in terms of raw image intensity in the neighborhood of the initial estimates. These centroid features are accurate to sub-pixel and can tolerate both slight defocus and smooth changes in illumination [125]. To increase robustness to outliers, we run the optimization for Eq. (3.6) iteratively, removing features whose reprojection error is more than 3.0 times the median.

### 3.6.4 Online geometric alignment

While the deterministic warp parameters need only be computed once for a given lens, we cannot apply the stochastic translations computed during calibration to a different sequence. Thus, when capturing images of a new scene, we must re-compute these translations.

In theory, it might be possible to identify key points and compute the best-fit translation. This would amount to redoing the optimization of Eq. (3.6) for each image independently, with all parameters except $\mathbf{T}$ fixed to the values computed offline. Unfortunately, feature localization can be unstable because different regions of the scene are defocused in different images. This makes sub-pixel feature estimation and alignment problematic at large apertures (see Fig. 3.1a, for example).

We deal with this issue by using Lucas-Kanade registration to compute the residual stochastic translations in an image-based fashion, assuming additive image noise [19, 30]. To avoid registration problems caused by defocus we (1) perform the alignment only between pairs of "adjacent" images (same focus and neighboring aperture, or vice versa) and (2) take into account only image patches with high frequency content. In particular, to align images taken at aperture settings $\alpha_i$, $\alpha_{i+1}$ and the same focus setting, we identify the patch of highest variance in the image taken at the maximum aperture, $\alpha_A$, and the same focus setting. Since this image produces maximum blur for defocused regions, patches with high frequency content in the images are guaranteed to contain high frequencies for any aperture.

## 3.7 Confocal Constancy Evaluation

Together, image alignment and relative exitance estimation allow us to establish a pixel-wise geometric and radiometric correspondence across all input images, *i.e.*, for all aperture and focus settings. Given a pixel $(x, y)$, we use this correspondence to assemble an $A \times F$ *aperture-focus image*, describing the pixel's intensity variations as a function of aperture and focus (Fig. 3.6a):

**The Aperture-Focus Image (AFI) of pixel** $(x, y)$

$$\mathbf{AFI}_{xy}(\alpha, f) = \frac{1}{\mathbf{R}_{xy}(\alpha, f)} \hat{\mathbf{I}}_{\alpha f}(x, y) \ , \tag{3.7}$$

where $\hat{\mathbf{I}}_{\alpha f}$ denotes the images after global lighting correction (Sec. 3.5) and geometric image alignment (Sec. 3.6).

AFIs are a rich source of information about whether or not a pixel is in focus at a particular focus setting $f$. We make this intuition concrete by developing two functionals that measure how well a pixel's AFI conforms to the confocal constancy property at $f$. Since we analyze the AFI of each pixel $(x, y)$ separately, we drop subscripts and use $\mathbf{AFI}(\alpha, f)$ to denote its AFI.

### 3.7.1 Direct Evaluation of Confocal Constancy

Confocal constancy tells us that when a pixel is in focus, its relative intensities across aperture should match the variation predicted by the relative exitance of the lens. Since Eq. (3.7) already corrects for these variations, confocal constancy at a hypothesis $\hat{f}$ implies constant intensity within column $\hat{f}$ of the AFI (Fig. 3.6b,c). Hence, to find the perfect focus setting we can simply find the column with minimum variance:

$$f^* = \arg\min_{\hat{f}} \operatorname{Var}\left[\left\{ \mathbf{AFI}(1, \hat{f}), \ldots, \mathbf{AFI}(A, \hat{f}) \right\}\right] . \tag{3.8}$$

To handle color images, we compute this cross-aperture variance for each RGB channel independently and then sum over channels.

The reason why the variance is higher at out-of-focus settings is that defocused pixels integrate regions of the scene surrounding the true surface point (Fig. 3.2b), which generally contain "texture" in the form of varying geometric structure or surface albedo. Hence, as with any method that does not use active illumination, the scene must contain sufficient spatial variation for this confocal constancy metric to be discriminative.

### 3.7.2 Evaluation by AFI Model-Fitting

A disadvantage of the previous method is that most of the AFI is ignored when testing a given focus hypothesis $\hat{f}$, since only one column of the AFI participates in the calculation of Eq. (3.8) (Fig. 3.6b). In reality, the 3D location of a scene point determines both the column of the AFI where confocal constancy holds as well as the degree of blur that occurs in the AFI's remaining, "out-of-focus" regions.[6] By taking these regions into account, we can create a focus detector with more resistance to noise and higher discriminative power.

In order to take into account both in- and out-of-focus regions of a pixel's AFI, we develop an idealized, parametric AFI model that generalizes confocal constancy. This model is controlled

[6]While not analyzed in the context of confocal constancy or the AFI, this is a key observation exploited by depth from defocus approaches [36, 42, 43, 92, 111, 120].

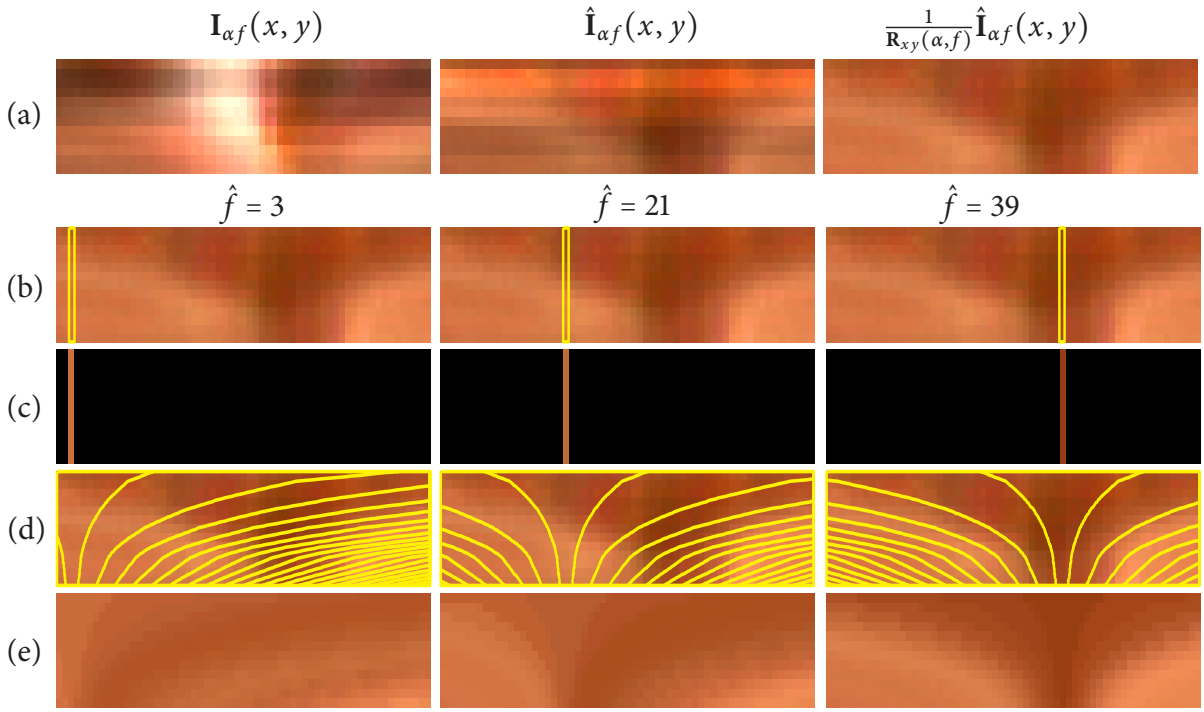



**Figure 3.6:** (a) The $A \times F$ measurements for the pixel shown in Fig. 3.1. *Left:* prior to image alignment. *Middle:* after image alignment. *Right:* after accounting for relative exitance (Eq. (3.7)). Note that the AFI's smooth structure is discernible only after both corrections. (b) Direct evaluation of confocal constancy for three focus hypotheses, $\hat{f} = 3, 21$ and 39. (c) Mean color of the corresponding AFI columns. (d) Boundaries of the equi-blur regions, superimposed over the AFI (for readability, only a third are shown). (e) Results of AFI model-fitting, with constant intensity in each equi-blur region, from the mean of the corresponding region in the AFI. Observe that for $\hat{f} = 39$ the model is in good agreement with the measured AFI ((a), rightmost).

by a single parameter—the focus hypothesis $\hat{f}$—and is fit directly to a pixel's AFI. The perfect focus setting is chosen to be the hypothesis that maximizes agreement with the AFI.

Our AFI model is based on two key observations. First, the AFI can be decomposed into a set of $F$ disjoint *equi-blur* regions that are completely determined by the focus hypothesis $\hat{f}$ (Fig. 3.6d). Second, under mild assumptions on scene radiance, the intensity within each equi-blur region will be constant when $\hat{f}$ is the correct hypothesis. These observations suggest that we can model the AFI as a set of $F$ constant-intensity regions whose spatial layout is determined by the focus hypothesis $\hat{f}$. Fitting this model to a pixel's AFI leads to a focus criterion that minimizes intensity variance in every equi-blur region (Fig. 3.6e):

$$f^* = \arg\min_{\hat{f}} \sum_{i=1}^{F} \left( w_i^{\hat{f}} \operatorname{Var}\left[ \left\{ \mathbf{AFI}(\alpha, f) \mid (\alpha, f) \in \mathcal{B}_i^{\hat{f}} \right\} \right] \right) , \tag{3.9}$$

where $\mathcal{B}_i^{\hat{f}}$ is the $i$-th equi-blur region for hypothesis $\hat{f}$, and $w_i^{\hat{f}}$ weighs the contribution of region

$\mathcal{B}_i^{\hat{f}}$. In our experiments, we set $w_i^{\hat{f}} = \text{area}(\mathcal{B}_i^{\hat{f}})$, which means that Eq. (3.9) reduces to computing the sum-of-squared error between the measured AFI and the AFI synthesized given each focus hypothesis. For color images, as in Eq. (3.8), we compute the focus criterion for each RGB channel independently and then sum over channels.

To implement Eq. (3.9) we must compute the equi-blur regions for a given focus hypothesis $\hat{f}$. Suppose that the hypothesis $\hat{f}$ is correct, and suppose that the current aperture and focus of the lens are $\alpha$ and $\hat{f}$, respectively, *i.e.*, a scene point $\hat{\mathbf{p}}$ is in perfect focus (Fig. 3.7a). Now consider "defocusing" the lens by changing its focus to $f$ (Fig. 3.7b). We can represent the blur associated with the pair $(\alpha, f)$ by a circular disc centered at point $\hat{\mathbf{p}}$ and parallel to the sensor plane. From similar triangles, the diameter of this disc is equal to

$$b_{\alpha f} = \frac{F}{\alpha} \frac{|\text{dist}(\hat{f}) - \text{dist}(f)|}{\text{dist}(f)} , \tag{3.10}$$

where $F$ is the focal length of the lens and $\text{dist}(\cdot)$ converts focus settings to distances from the aperture.[7] Our representation of this function assumes that the focal surfaces are fronto-parallel planes [105].

Given a focus hypothesis $\hat{f}$, Eq. (3.10) assigns a "blur diameter" to each point $(\alpha, f)$ in the AFI and induces a set of nested, wedge-shaped curves of equal blur diameter (Figs. 3.6d and 3.7). We quantize the possible blur diameters into $F$ bins associated with the widest-aperture settings, *i.e.*, $(\alpha_A, f_1), \ldots, (\alpha_A, f_F)$, which partitions the AFI into $F$ equi-blur regions, one per bin.

Eq. (3.10) fully specifies our AFI model, and we have found that this model matches the observed pixel variations quite well in practice (Fig. 3.6e). It is important, however, to note that this model is approximate. In particular, we have implicitly assumed that once relative exitance and geometric distortion have been factored out (Secs. 3.5–3.6), the equi-blur regions of the AFI are well-approximated by the equi-blur regions predicted by the thin-lens model [16, 105]. Then, the intensity at two positions in an equi-blur region will be constant under the following conditions: (i) the largest aperture subtends a small solid angle from all scene points, (ii) outgoing radiance for all scene points contributing to a defocused pixel remains constant within the cone of the largest aperture, and (iii) depth variations for such scene points do not significantly affect the defocus integral. See Appendix B for a formal analysis.

[7] To calibrate the function $\text{dist}(\cdot)$, we used the same calibration pattern as in Sec. 3.6, mounted on a translation stage parallel to the optical axis. For various stage positions spanning the workspace, we used the camera's autofocus feature and measured the corresponding focus setting using a printed ruler mounted on the lens. We related stage positions to absolute distances using a FaroArm Gold 3D touch probe, whose single-point accuracy was ±0.05 mm.

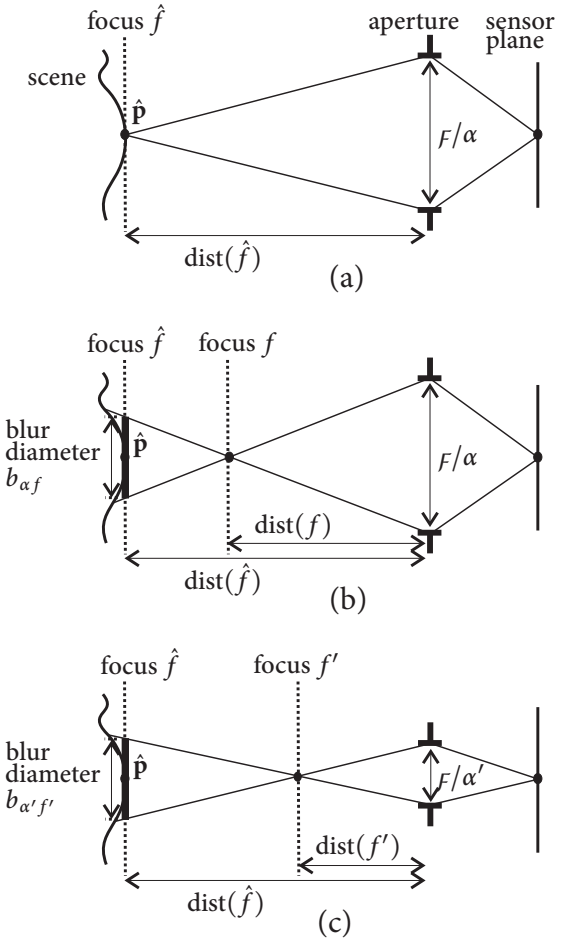



**Figure 3.7:** Quantifying the blur due to an out-of-focus setting. (a) At focus setting $\hat{f}$, scene point $\hat{\mathbf{p}}$ is in perfect focus. The aperture's effective diameter can be expressed in terms of its f-stop value $\alpha$ and the focal length $F$. (b) For an out-of-focus setting $f$, we can use Eq. (3.10) to compute the effective blur diameter, $b_{\alpha f}$. (c) A second aperture-focus combination with the same blur diameter, $b_{\alpha' f'} = b_{\alpha f}$. In our AFI model, $(\alpha, f)$ and $(\alpha', f')$ belong to the same equi-blur region.

## 3.8 Experimental Results

To test our approach we used two setups representing different grades of camera equipment. Our first setup was designed to test the limits of pixel-level reconstruction accuracy in a high-resolution setting, by using professional-quality camera with a wide-aperture lens. In the second setup, we reproduced our approach with older and low-quality equipment, using one of earliest digital SLR cameras, with a low-quality zoom lens.

For the first setup, we used two different digital SLR cameras, the 16 MP Canon EOS-1Ds Mark II (BOX dataset), and the 12 MP Canon EOS-1Ds (HAIR and PLASTIC datasets). For both cameras we used the same wide-aperture, fixed focal length lens (Canon EF85mm f1.2L). The lens aperture was under computer control and its focal setting was adjusted manually using

a printed ruler on the body of the lens. We operated the cameras at their highest resolution, capturing 4992 × 3328-pixel and 4604 × 2704-pixel images respectively in RAW 12-bit mode. Each image was demosaiced using Canon software and linearized using the algorithm in [31]. We used $A = 13$ apertures ranging from f/1.2 to f/16, and $F = 61$ focal settings spanning a workspace that was 17 cm in depth and 1.2 m away from the camera. Successive focal settings therefore corresponded to a depth difference of approximately 2.8 mm. We mounted the camera on an optical table in order to allow precise ground-truth measurements and to minimize external vibrations.

For the second setup, we used a 6 MP Canon 10D camera (Teddy dataset) with a low-quality zoom-lens (Canon EF24-85mm f3.5-4.5). Again, we operated the camera in RAW mode at its highest resolution, which here was 3072 × 2048. Unique to this setup, we manipulated focal setting using a computer-controlled stepping motor to drive the lens focusing ring mechanically [4]. We used $A = 11$ apertures ranging from f/3.5 to f/16, and $F = 41$ focal settings spanning a workspace that was 1.0 m in depth and 0.5 m away from the camera. Because this lens has a smaller maximum aperture, the depth resolution was significantly lower, and the distance between successive focal settings was over 8 mm at the near end of the workspace.[8]

To enable the construction of aperture-focus images, we first computed the relative exitance of the lens (Sec. 3.5) and then performed offline geometric calibration (Sec. 3.6). For the first setup, our geometric distortion model was able to align the calibration images with an accuracy of approximately 0.15 pixels, as estimated from centroids of dot features (Fig. 3.5c). The accuracy of online alignment was about 0.4 pixels, *i.e.*, worse than during offline calibration but well below one pixel. This penalty is expected since we use smaller regions of the scene for online alignment, and since we align the image sequence in an incremental pairwise fashion, to avoid alignment problems with severely defocused image regions (see Sec. 3.6.4). Calibration accuracy for the second setup was similar.

While the computation required by confocal stereo is simple and linear in the total number of pixels and focus hypotheses, the size of the datasets make memory size and disk speed the main computational bottlenecks. In our experiments, image capture took an average of two seconds per frame, demosaicking one minute per frame, and alignment and further preprocessing about three minutes per frame. For a 128 × 128 pixel patch, a Matlab implementation of AFI model-fitting took about 250 s using 13 × 61 images, compared with 10 s for a depth from focus method that uses 1 × 61 images.

[8] For additional results, see `http://www.cs.toronto.edu/~hasinoff/confocal`.

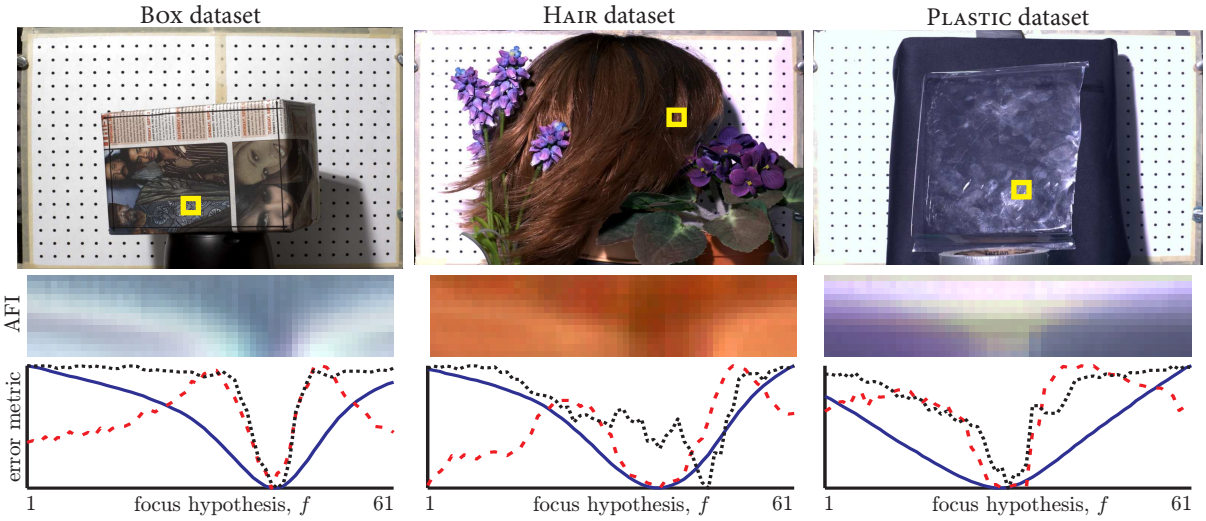



**Figure 3.8:** Behavior of focus criteria for a specific pixel (highlighted square) in three test datasets. The dashed graph is for direct confocal constancy (Eq. (3.8)), solid is for AFI model-fitting (Eq. (3.9)), and the dotted graph is for $3 \times 3$ variance (DFF). While all three criteria often have corresponding local minima near the perfect focus setting, AFI model-fitting varies much more smoothly and exhibits no spurious local minima in these examples. For the middle example, which considers the same pixel shown in Fig. 3.1, the global minimum for variance is at an incorrect focus setting. This is because the pixel lies on a strand of hair only 1–2 pixels wide, beyond the resolving power of variance calculations. The graphs for each focus criterion are shown with relative scaling.

**Quantitative evaluation: Box dataset.** To quantify reconstruction accuracy, we used a tilted planar scene consisting of a box wrapped in newsprint (Fig. 3.8, left). The plane of the box was measured using a FaroArm Gold 3D touch probe, as employed in Sec. 3.7.2, whose single-point accuracy was ±0.05 mm in the camera's workspace. To relate probe coordinates to coordinates in the camera's reference frame we used the Camera Calibration Toolbox for Matlab [23] along with further correspondences between image features and 3D coordinates measured by the probe.

We computed a depth map of the scene for three focus criteria: direct confocal constancy (Eq. (3.8)), AFI model-fitting (Eq. (3.9)), and a depth from focus (DFF) method, applied to the widest-aperture images, that chooses the focus setting with the highest variance in a $3 \times 3$ window centered at each pixel, summed over RGB color channels. The planar shape of the scene and its detailed texture can be thought of as a best-case scenario for such window-based approaches. The plane's footprint contained 2.8 million pixels, yielding an equal number of 3D measurements.

As Table 3.1 shows, all three methods performed quite well, with accuracies of 0.36–0.49 % of the object-to-camera distance. This performance is on par with previous quantitative studies

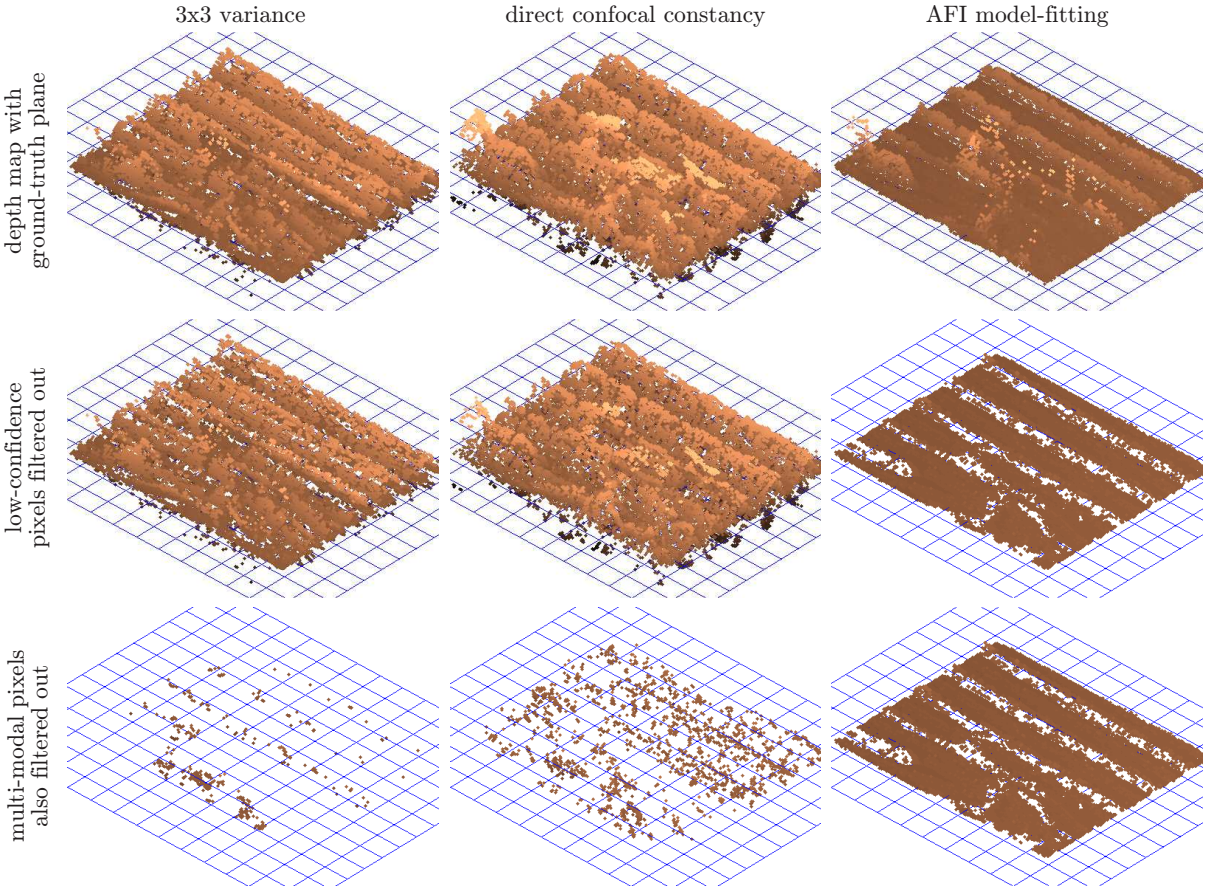



**Figure 3.9:** Visualizing the accuracy of reconstruction and outlier detection for the Box dataset. *Top row:* For all three focus criteria, we show depth maps for a 200 × 200 region from the center of the box (see Fig. 3.10). The depth maps are rendered as 3D point clouds where intensity encodes depth, and with the ground-truth plane shown overlaid as a 3D mesh. *Middle row:* We compute confidence for each pixel as the second derivative at the minimum of the focus criterion. For comparison across different focus criteria, we fixed the threshold for AFI model-fitting, and adjusted the thresholds so that the other two criteria reject the same number of outliers. While this significantly helps reject outliers for AFI model-fitting, for the other criteria, which are typically multi-modal, this strategy is much less effective. *Bottom row:* Subsequently filtering out pixels with multiple modes has little effect on AFI model-fitting, which is nearly always uni-modal, but removes almost all pixels for the other criteria.

**Table 3.1:** Ground-truth accuracy results. All distances were measured relative to the ground-truth plane, and the inlier threshold was set to 11 mm. We also express the RMS error as a percentage of the mean camera-to-scene distance of 1025 mm.

| | median abs. dist. (mm) | inlier RMS dist. (mm) | % inliers | RMS % dist. to camera |
|---|---|---|---|---|
| 3 × 3 spatial variance (DFF) | 2.16 | 3.79 | 80 | 0.374 |
| confocal constancy evaluation | 3.47 | 4.99 | 57 | 0.487 |
| AFI model-fitting | 2.14 | 3.69 | 91 | 0.356 |

[120, 132], although few results with real images have been reported in the passive depth from focus literature. Significantly, AFI model-fitting slightly outperforms spatial variance (DFF) in both accuracy and number of outliers even though its focus computations are performed entirely at the pixel level and, hence, are of much higher resolution. Qualitatively, this behavior is confirmed by considering all three criteria for specific pixels (Fig. 3.8) and for an image patch (Figs. 3.9 and 3.10).

Note that it is also possible to detect outlier pixels where the focus criterion is uninformative (*e.g.*, when the AFI is nearly constant due to lack of texture) by using a confidence measure or by processing the AFI further. We have experimented with a simple confidence measure computed as the second derivative at the minimum of the focus criterion[9]. As shown in Fig. 3.9, filtering out low-confidence pixels for AFI model-fitting leads to a sparser depth map that suppresses noisy pixels, but for the other focus criteria, where most pixels have multiple modes, such filtering is far less beneficial. This suggests that AFI model-fitting is a more discriminative focus criterion, because it produces fewer modes that are both sharply peaked and incorrect.

As a final experiment with this dataset, we investigated how AFI model-fitting degrades when a reduced number of apertures is used (*i.e.*, for AFIs of size $A' \times F$ with $A' < A$). Our results suggest that using only five or six apertures causes little reduction in quality (Fig. 3.11).

**HAIR dataset.** Our second test scene was a wig with a messy hairstyle, approximately 25 cm tall, surrounded by several artificial plants (Figs. 3.1 and 3.8, middle). Reconstruction results for this scene (Fig. 3.12) show that our confocal constancy criteria lead to very detailed depth maps, at the resolution of individual strands of hair, despite the scene's complex geometry and despite the fact that depths can vary greatly within small image neighborhoods (*e.g.*, toward the silhouette of the wig). By comparison, the $3 \times 3$ variance operator produces uniformly-lower resolution results, and generates smooth "halos" around narrow geometric structures like individual strands of hair. In many cases, these "halos" are larger than the width of the spatial operator, as blurring causes distant points to influence the results.

In low-texture regions, such as the cloth flower petals and leaves, fitting a model to the entire AFI allows us to exploit defocused texture from nearby scene points. Window-based methods like variance, however, generally yield even better results in such regions, because they propagate focus information from nearby texture more directly, by implicitly assuming a smooth scene

[9] In practice, since computing second derivatives directly can be noisy, we compute the width of the valley that contains the minimum, at a level 10 % above the minimum. For AFI model-fitting across all datasets, we reject pixels whose width exceeds 14 focus settings. Small adjustments to this threshold do not change the results significantly.

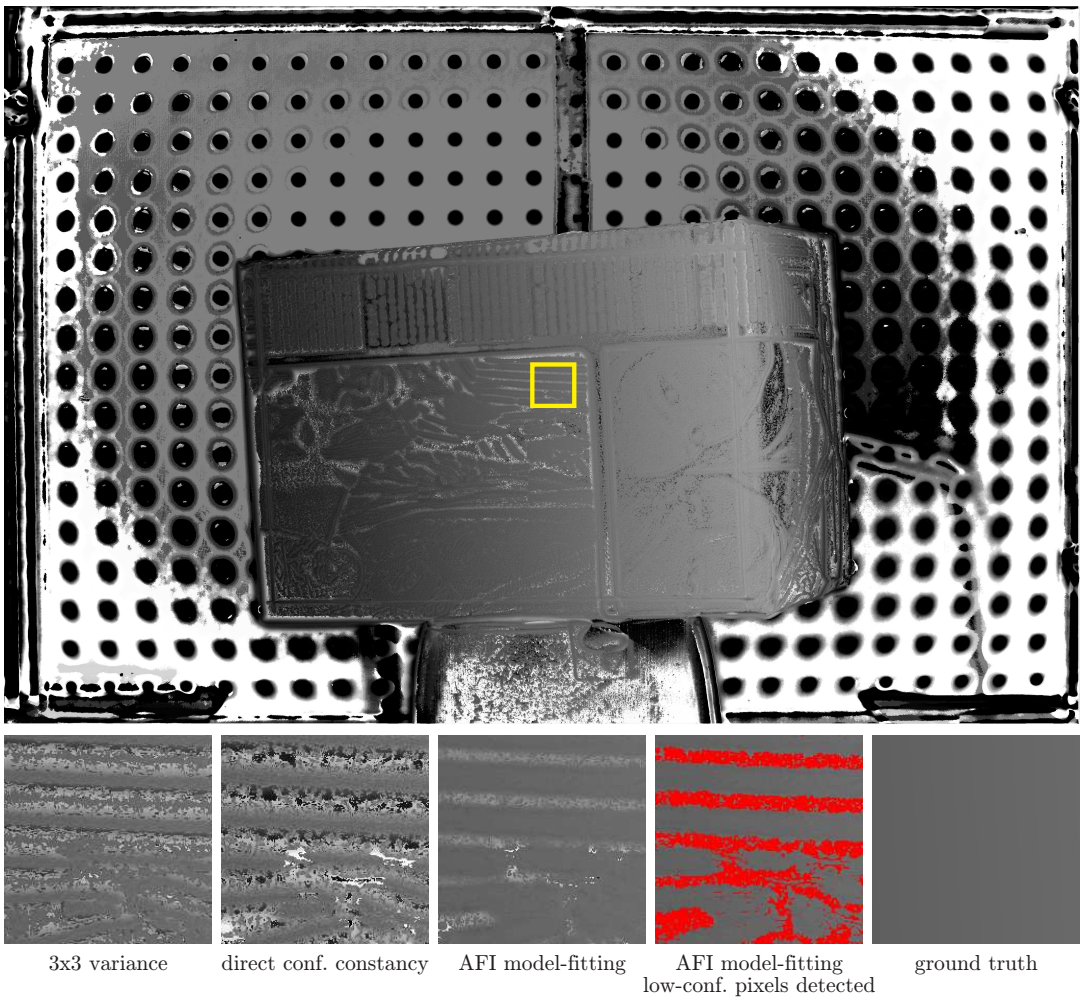



**Figure 3.10:** *Top:* Depth map for the Box dataset using AFI model-fitting. *Bottom:* Close-up depth maps for the highlighted region corresponding to Fig. 3.9, computed using three focus criteria.

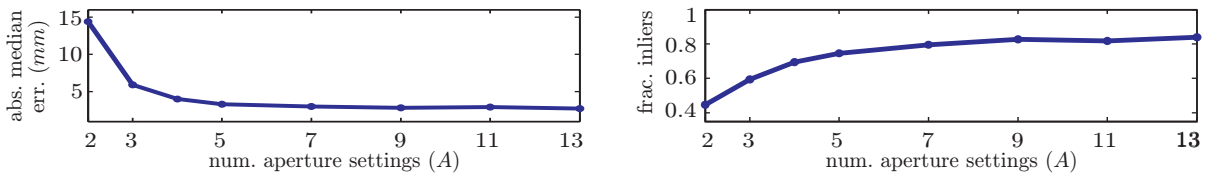



**Figure 3.11:** AFI model-fitting error and inlier fraction as a function of the number of aperture settings (Box dataset, inlier threshold = 11 mm).

geometry. Like all focus measures, those based on confocal constancy are uninformative in extremely untextured regions, *i.e.*, when the AFI is constant. However, by using the proposed confidence measure, we can detect many of these low-texture pixels (Figs. 3.12 and 3.15). To better visualize the result of filtering out these pixels, we replace them using a simple variant of PDE-based inpainting [21].

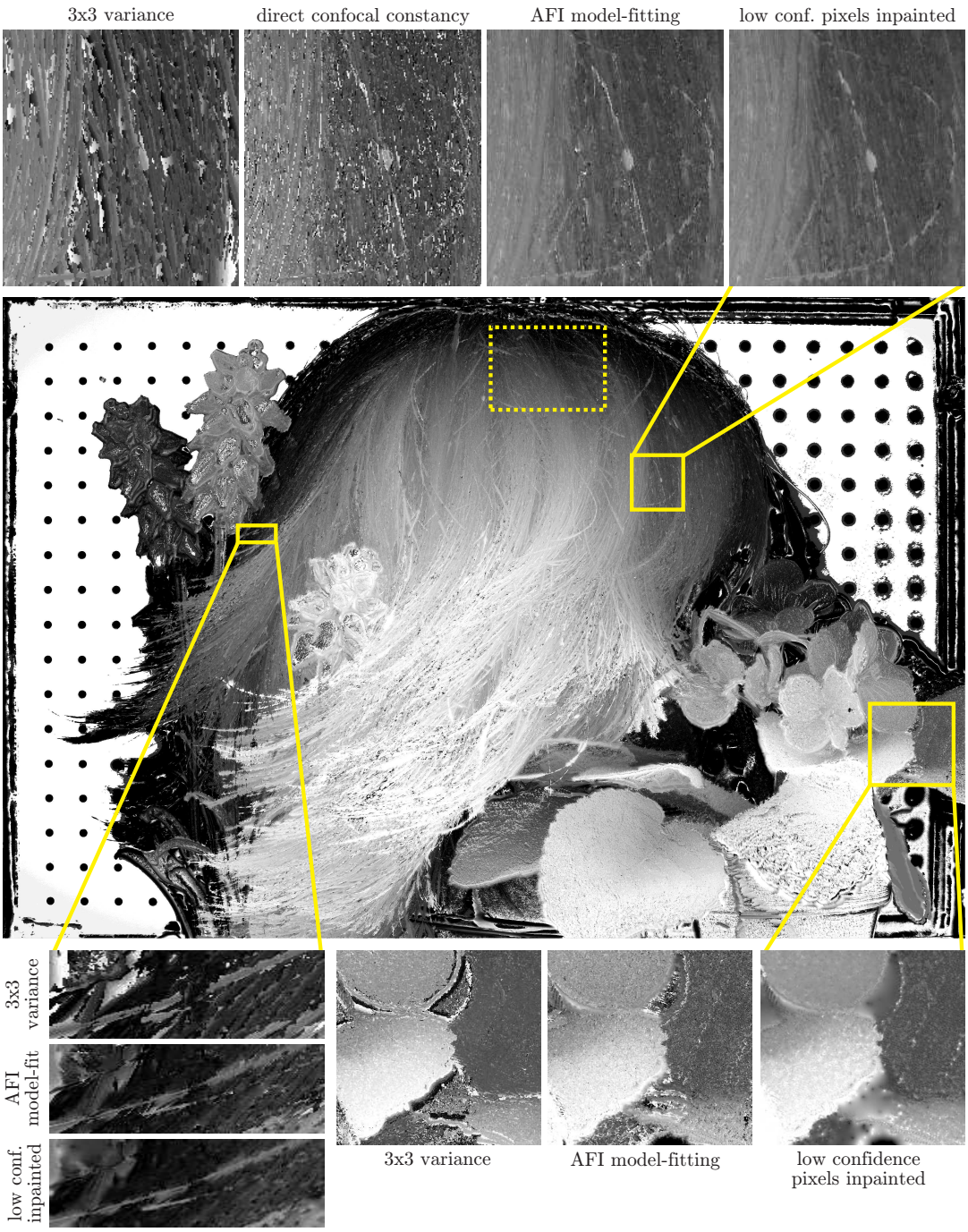



**Figure 3.12:** *Center:* Depth map for the HAIR dataset using AFI model-fitting. *Top:* The AFI-based depth map resolves several distinctive foreground strands of hair. We also show the result of detecting low-confidence pixels from AFI model-fitting and replacing them using PDE-based inpainting [21] (see Fig. 3.15), which suppresses noise but preserves fine detail. Direct evaluation of confocal constancy is also sharp but much noisier, making structure difficult to discern. By contrast, $3\times3$ variance (DFF) exhibits thick "halo" artifacts and fails to detect most of the foreground strands (see also Fig. 3.8). *Bottom right:* DFF yields somewhat smoother depths for the low-texture leaves, but exhibits inaccurate halo artifacts at depth discontinuities. *Bottom left:* Unlike DFF, AFI model-fitting resolves structure amid significant depth discontinuities.

**Plastic dataset.** Our third test scene was a rigid, near-planar piece of transparent plastic, formerly used as packaging material, which was covered with dirt, scratches, and fingerprints (Fig. 3.8, right). This plastic object was placed in front of a dark background and lit obliquely to enhance the contrast of its limited surface texture. Reconstruction results for this scene (Figs. 3.13–3.14) illustrate that at high resolution, even transparent objects may have enough fine-scale surface texture to be reconstructed using focus- or defocus-based techniques. In general, wider baseline methods like standard stereo cannot exploit such surface texture easily because textured objects behind the transparent surface may interfere with matching.

Despite the scene's relatively low texture, AFI model-fitting still recovers the large-scale planar geometry of the scene, albeit with significant outliers (Fig. 3.14). By comparison, the $3 \times 3$ variance operator recovers a depth map with fewer outliers, which is expected since window-based approaches are well suited to reconstruction of near-planar scenes. As in the previous dataset, most of the AFI outliers can be attributed to low-confidence pixels and are readily filtered out (Fig. 3.15).

**Teddy dataset.** Our final test scene, captured using low-quality camera equipment, consists of a teddy bear with coarse fur, seated in front of a hat and several cushions, with a variety of ropes in the foreground (Fig. 3.16). Since little of this scene is composed of the fine pixel-level texture found in previous scenes, this final dataset provides an additional test for low-texture areas.

We had no special difficulty applying our method for this new setup, and even with a lower-quality lens we obtained a similar level of accuracy with our radiometric and geometric calibration model. As shown in Fig. 3.16, the results are qualitatively comparable to depth recovery for the low-texture objects in previous datasets. The large-scale geometry of the scene is clearly recovered, and many of the outliers produced by our pixel-level AFI model-fitting method can be identified as well.

**Online alignment.** To qualitatively assess the effect of online alignment, which accounts for both stochastic sub-pixel camera motion (Sec. 3.6.4) as well as temporal variations in lighting intensity (Sec. 3.5), we compared the depth maps produced using AFI model-fitting (Eq. (3.9)) with and without this alignment step (Fig. 3.15a,b). Our results show that online alignment leads to noise reduction for low-texture, dark, or other noisy pixels (*e.g.*, due to color demosaicking), but does not resolve significant additional detail. This also suggests that any further improvements to geometric calibration might lead to only slight gains.

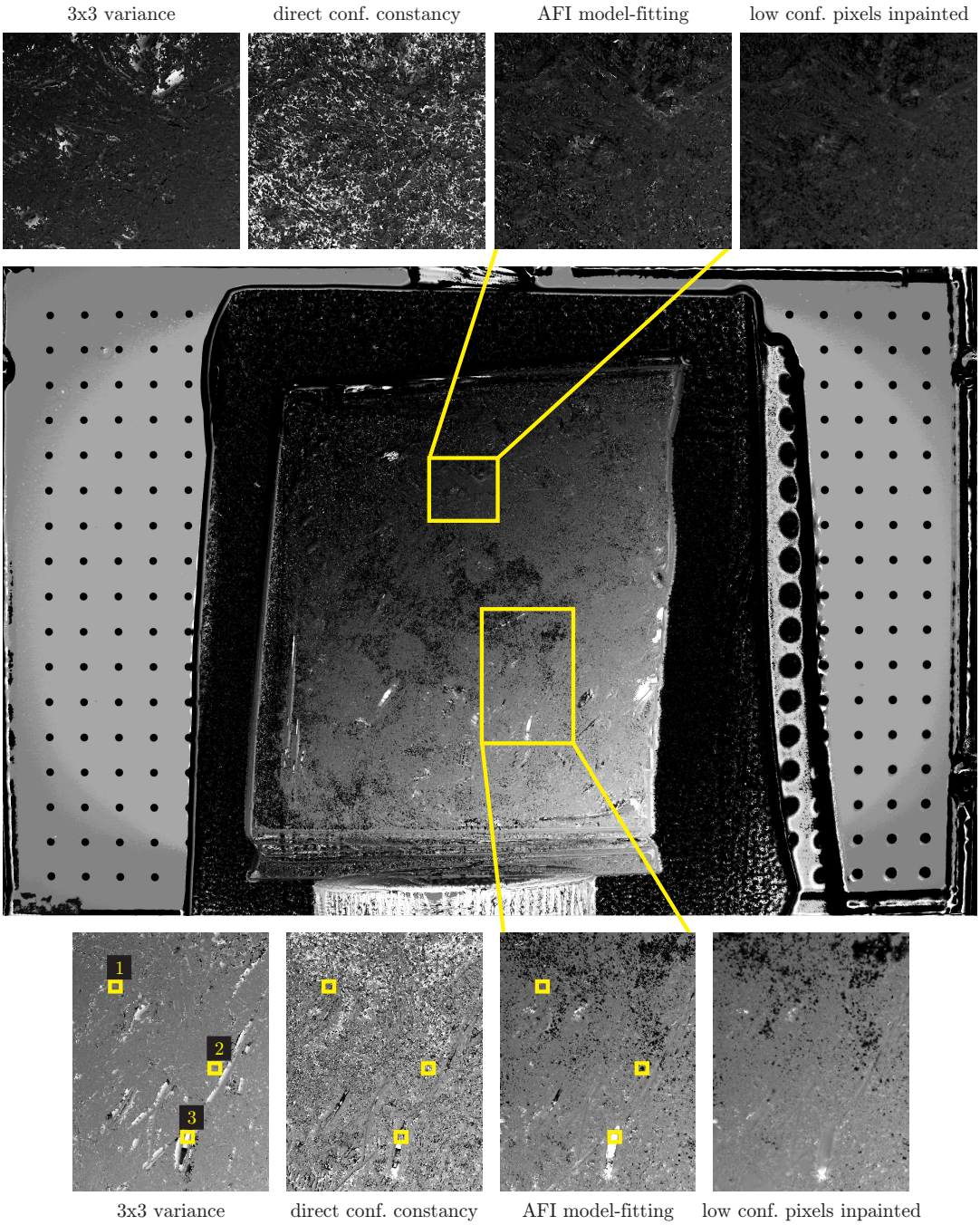



**Figure 3.13:** *Center:* Depth map for the PLASTIC dataset using AFI model-fitting. *Top:* Close-up depth maps for the highlighted region, computed using three focus criteria. While 3×3 variance (DFF) yields the smoothest depth map overall for the transparent surface, there are still a significant number of outliers. Direct evaluation of confocal constancy, is extremely noisy for this dataset, but AFI model-fitting recovers the large-scale smooth geometry. *Bottom:* Similar results for another highlighted region of the surface, but with relatively more outliers for AFI model-fitting. While AFI model-fitting produces more outliers overall than DFF for this dataset, many of these outliers can be detected and replaced using inpainting. Focus criteria for the three highlighted pixels are shown in Fig. 3.14.

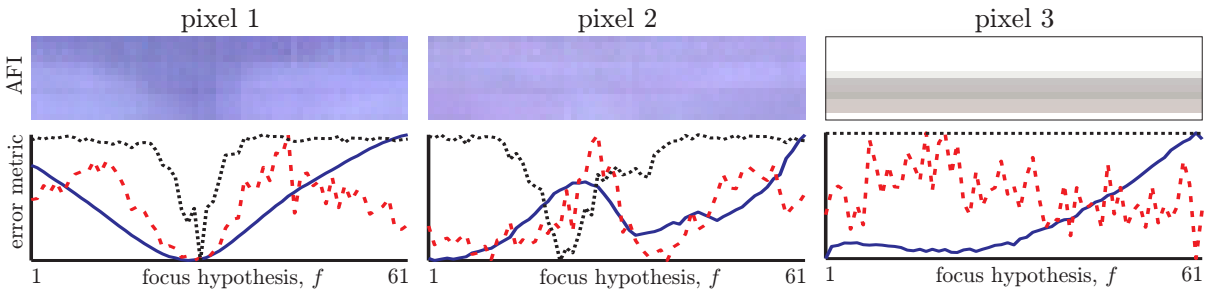



**Figure 3.14:** Failure examples. *Left to right:* Behavior of the three focus criteria in Fig. 3.13 for three highlighted pixels. The dashed graph is for direct confocal constancy (Eq. (3.8)), solid is for AFI model-fitting (Eq. (3.9)), and the dotted graph is for $3 \times 3$ variance (DFF). For pixel 1 all minima coincide. Lack of structure in pixel 2 produces multiple local minima for the AFI model-fitting metric; only DFF provides an accurate depth estimate. Pixel 3 and its neighborhood are corrupted by saturation, so no criterion gives meaningful results. Depth estimates at pixel 2 and 3 would have been rejected by our confidence criterion.

Four observations can be made from our experiments. First, we have validated the ability of confocal stereo to estimate depths for fine pixel-level geometric structures. Second, the radiometric calibration and image alignment method we use are sufficient to allow us to extract depth maps with very high resolution cameras and wide-aperture lenses. Third, our method can still be applied successfully in a low-resolution setting, using low-quality equipment. Fourth, although the AFI is uninformative in completely untextured regions, we have shown that a simple confidence metric can help identify such pixels, and that AFI model-fitting can exploit defocused texture from nearby scene points to provide useful depth estimates even in regions with relatively low texture.

## 3.9 Discussion and Limitations

The extreme locality of shape computations derived from aperture-focus images is both a key advantage and a major limitation of the current approach. While we have shown that processing a pixel's AFI leads to highly detailed reconstructions, this locality does not yet provide the means to handle large untextured regions [38, 114] or to reason about global scene geometry and occlusion [16, 42, 99].

Untextured regions of the scene are clearly problematic since they lead to near-constant and uninformative AFIs. The necessary conditions for resolving scene structure, however, are even more stringent because a fronto-parallel plane colored with a linear gradient can also produce constant AFIs.[10] To handle these cases, we are exploring the possibility of analyzing AFIs at

[10] This follows from the work of Favaro, *et al.* [38] who established that non-zero second-order albedo gradients

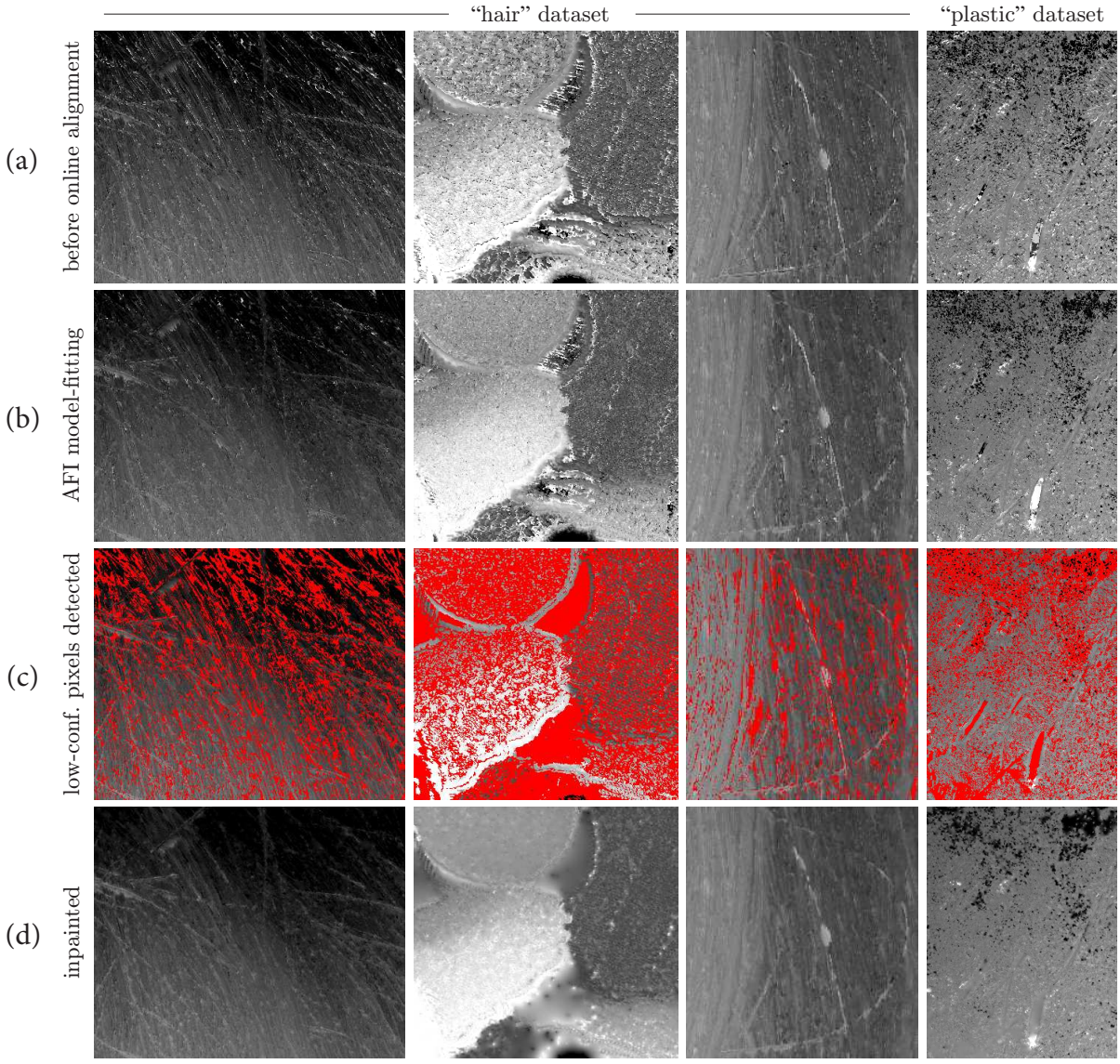



**Figure 3.15:** (a)–(b) Improvement of AFI model-fitting due to online alignment, accounting for stochastic sub-pixel camera motion and temporal variations in lighting intensity. (b) Online alignment leads to a reduction in noisy pixels and yields smoother depth maps for low-textured regions, but does not resolve significantly more detail in our examples. (c) Low-confidence pixels for the AFI model-fitting criterion, highlighted in red, are pixels where the second derivative at the minimum is below the same threshold used for AFI model-fitting in Fig. 3.9. (d) Low confidence pixels filled using PDE-based inpainting [21]. By comparison to (b), we see that many outliers have been filtered, and that the detailed scene geometry has been preserved. The close-up depth maps correspond to regions highlighted in Figs. 3.12–3.13.

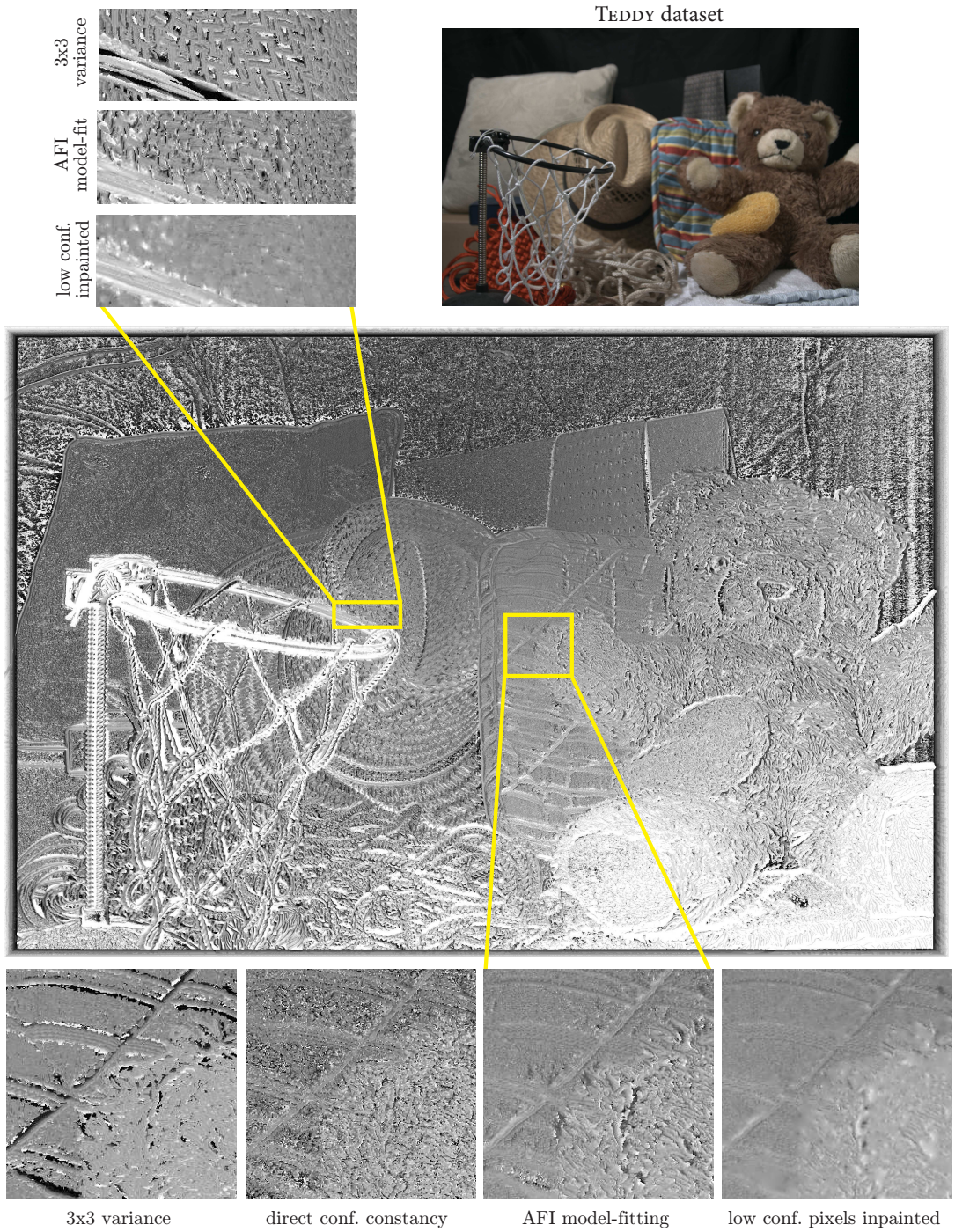



**Figure 3.16:** *Top right:* Sample widest-aperture f/3.5 input photo of the Teddy dataset. *Center:* Depth map using AFI model-fitting. *Top left:* Close-up depth maps for the highlighted region, comparing 3×3 variance (DFF) and AFI model-fitting, with and without inpainting of the detected outliers. Like the Plastic dataset shown in Fig. 3.13, outliers are significant for low-texture regions. While window-based DFF leads to generally smoother depths, AFI model-fitting provides the ability to distinguish outliers. *Bottom:* Similar effects can be seen for the bear's paw, just in front of low-texture cushion.

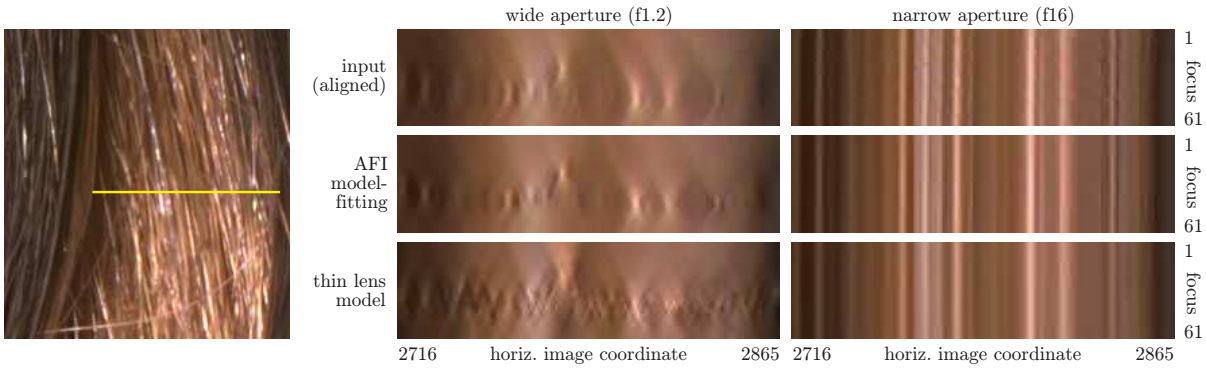



**Figure 3.17:** AFI model-fitting *vs*. the thin lens model. *Left:* Narrow-aperture image region from the HAIR dataset, corresponding to Fig. 3.12, top. *Right:* For two aperture settings, we show the cross-focus appearance variation of the highlighted horizontal segment: (i) for the aligned input images, (ii) re-synthesized using AFI model-fitting, and (iii) re-synthesized using the thin lens model. To resynthesize the input images we used the depths and colors predicted by AFI model-fitting. At wide apertures, AFI model-fitting much better reproduces the input, but at the narrowest aperture both methods are identical.

multiple levels of detail and analyzing the AFIs of multiple pixels simultaneously. The goal of this general approach is to enforce geometric smoothness only when required by the absence of structure in the AFIs of individual pixels.

Although not motivated by the optics, it is also possible to apply Markov random field (MRF) optimization, *e.g*., [24], to the output of our per-pixel analysis, since Eqs. (3.8) and (3.9) effectively define "data terms" measuring the level of inconsistency for each depth hypothesis. Such an approach would bias the reconstruction toward piecewise-smooth depths, albeit without exploiting the structure of defocus over spatial neighborhoods. To emphasize our ability to reconstruct pixel-level depth we have not taken this approach, but have instead restricted ourselves to a greedy per-pixel analysis.

Since AFI's equi-blur regions are derived from the thin lens model, it is interesting to compare our AFI model's ability to account for the input images, compared to the pure thin lens model. In this respect, the fitted AFIs are much better at capturing the spatial and cross-focus appearance variations (Fig. 3.17). Intuitively, our AFI model is less constrained than the thin lens model, because it depends on $F$ color parameters per pixel (one for each equi-blur region), instead of just one. Furthermore, these results suggest that lens defocus may be poorly described by simple analytic point-spread functions as in existing methods, and that more expressive models based on the structure of the AFI may be more useful in fully accounting for defocus.

Finally, as a pixel-level method, confocal stereo exhibits better behavior near occlusion bound-

are a necessary condition for resolving the structure of a smooth scene.

aries compared to standard defocus-based techniques that require integration over spatial windows. Nevertheless, confocal constancy does not hold exactly for pixels that are both near an occlusion boundary and correspond to the occluded surface because the assumption of a fully-visible aperture breaks down. To this end, we are investigating more explicit methods for occlusion modeling [16, 42], as well as the use of a space-sweep approach to account for these occlusions, analogous to voxel-based stereo [68].

## Summary

The key idea of our approach is the introduction of the aperture-focus image, which serves as an important primitive for depth computation at high resolutions. We showed how each pixel can be analyzed in terms of its AFI, and how this analysis led to a simple method for estimating depth at each pixel individually. Our results show that we can compute 3D shape for very complex scenes, recovering fine, pixel-level structure at high resolution. We also demonstrated ground truth results for a simple scene that compares favorably to previous methods, despite the extreme locality of confocal stereo computations.

Although shape recovery is our primary motivation, we have also shown how, by computing an empirical model of a lens, we can achieve geometric and radiometric image alignment that closely matches the behavior and capabilities of high-end consumer lenses and imaging sensors. In this direction, we are interested in exploiting the typically unnoticed stochastic, sub-pixel distortions in SLR cameras in order to achieve super-resolution [90], as well as for other applications.

# Chapter 4

# Layer-Based Restoration for Multiple-Aperture Photography

> There are two kinds of light—the glow that illuminates, and the glare that obscures.
>
> James Thurber (1894–1961)

In this chapter we present *multiple-aperture photography*, a new method for analyzing sets of images captured with different aperture settings, with all other camera parameters fixed. We show that by casting the problem in an image restoration framework, we can simultaneously account for defocus, high dynamic range exposure (HDR), and noise, all of which are confounded according to aperture. Our formulation is based on a layered decomposition of the scene that models occlusion effects in detail. Recovering such a scene representation allows us to adjust the camera parameters in post-capture, to achieve changes in focus setting or depth of field—with all results available in HDR. Our method is designed to work with very few input images: we demonstrate results from real sequences obtained using the three-image "aperture bracketing" mode found on consumer digital SLR cameras.

## 4.1 Introduction

Typical cameras have three major controls—aperture, shutter speed, and focus. Together, aperture and shutter speed determine the total amount of light incident on the sensor (*i.e.*, exposure), whereas aperture and focus determine the extent of the scene that is in focus (and the degree of out-of-focus blur). Although these controls offer flexibility to the photographer, once an image has been captured, these settings cannot be altered.

Recent computational photography methods aim to free the photographer from this choice by collecting several controlled images [10, 34, 78], or using specialized optics [61, 85]. For example, high dynamic range (HDR) photography involves fusing images taken with varying shutter speed, to recover detail over a wider range of exposures than can be achieved in a single photo [5, 78].

In this chapter we show that flexibility can be greatly increased through multiple-aperture photography, *i.e.*, by collecting several images of the scene with all settings except aperture fixed (Fig. 4.1). In particular, our method is designed to work with very few input images, including the three-image "aperture bracketing" mode found on most consumer digital SLR cameras. Multiple-aperture photography takes advantage of the fact that by controlling aperture we simultaneously modify the exposure and defocus of the scene. To our knowledge, defocus has not previously been considered in the context of widely-ranging exposures.

We show that by inverting the image formation in the input photos, we can decouple all three controls—aperture, focus, and exposure—thereby allowing complete freedom in post-capture, *i.e.*, we can resynthesize HDR images for any user-specified focus position or aperture setting. While this is the major strength of our technique, it also presents a significant technical challenge. To address this challenge, we pose the problem in an image restoration framework, connecting the radiometric effects of the lens, the depth and radiance of the scene, and the defocus induced by aperture.

The key to the success of our approach is formulating an image formation model that accurately accounts for the input images, and allows the resulting image restoration problem to be inverted in a tractable way, with gradients that can be computed analytically. By applying the image formation model in the forward direction we can resynthesize images with arbitrary camera settings, and even extrapolate beyond the settings of the input.

In our formulation, the scene is represented in layered form, but we take care to model occlusion effects at defocused layer boundaries [16] in a physically meaningful way. Though several depth-from-defocus methods have previously addressed such occlusion, these methods have been limited by computational inefficiency [42], a restrictive occlusion model [22], or the assumption that the scene is composed of two surfaces [22, 42, 77]. By comparison, our approach can handle an arbitrary number of layers, and incorporates an approximation that is effective and efficient to compute. Like McGuire, *et al.* [77], we formulate our image formation model in terms of image compositing [104], however our analysis is not limited to a two-layer scene or input photos with special focus settings.

Our work is also closely related to depth-from-defocus methods based on image restoration,

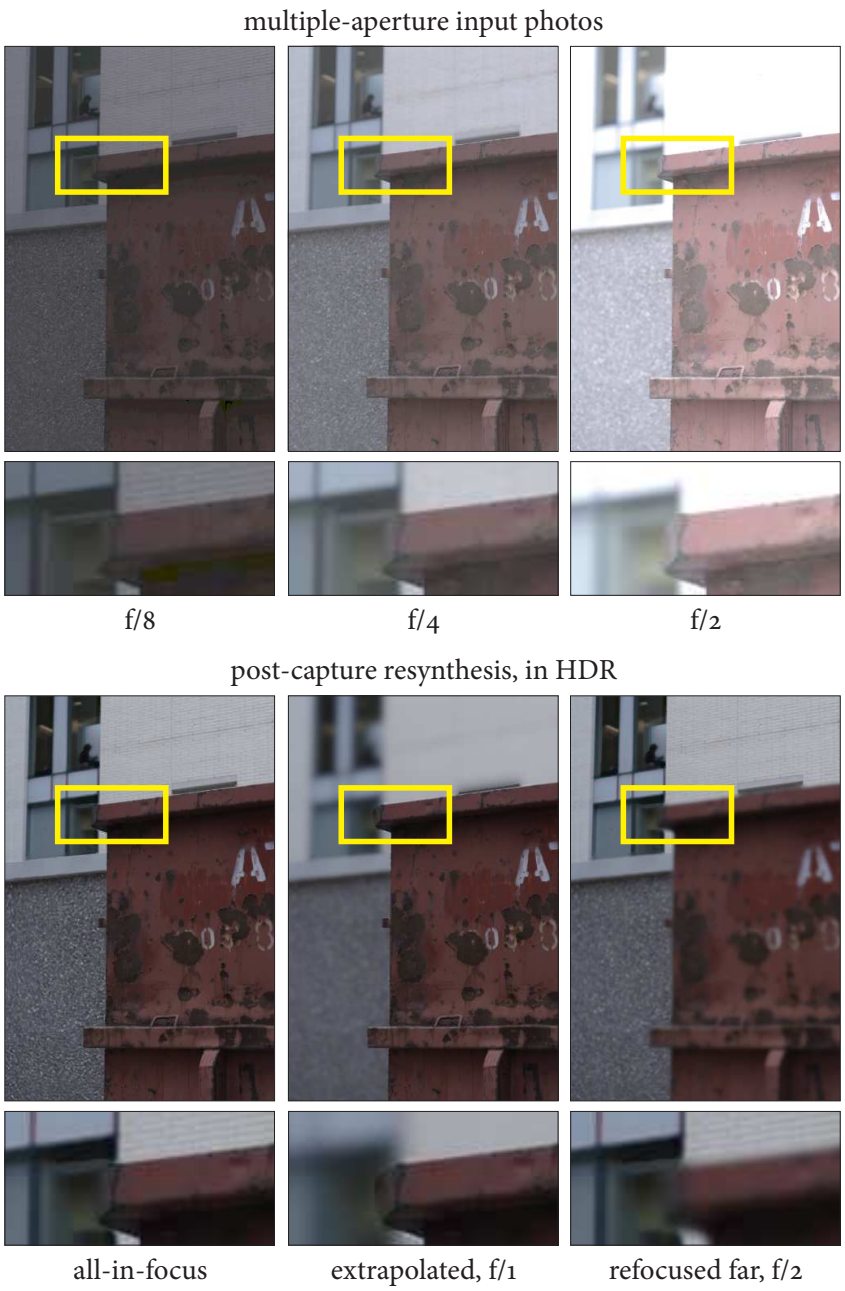



**Figure 4.1:** Photography with varying apertures. *Top:* Input photographs for the Dumpster dataset, obtained by varying aperture setting only. Without the strong gamma correction we apply for display ($\gamma = 3$), these images would appear extremely dark or bright, since they span a wide exposure range. Note that aperture affects both exposure and defocus. *Bottom:* Examples of post-capture resynthesis, shown in high dynamic range (HDR) with tone-mapping. Left-to-right: the all-in-focus image, an extrapolated aperture (f/1), and refocusing on the background (f/2).

that recover an all-in-focus representation of the scene [42, 62, 95, 107]. Although the output of these methods theoretically permits post-capture refocusing and aperture control, most of these methods assume an additive, transparent image formation model [62, 95, 107] which causes seri-

ous artifacts at depth discontinuities, due to the lack of occlusion modeling. Similarly, defocus-based techniques specifically designed to allow refocusing rely on inverse filtering with local windows [14, 29], and do not model occlusion either. Importantly, none of these methods are designed to handle the large exposure differences found in multiple-aperture photography.

Our work has four main contributions. First, we introduce multiple-aperture photography as a way to decouple exposure and defocus from a sequence of images. Second, we propose a layered image formation model that is efficient to evaluate, and enables accurate resynthesis by accounting for occlusion at defocused boundaries. Third, we show that this formulation is specifically designed for an objective function that can be practicably optimized within a standard restoration framework. Fourth, as our experimental results demonstrate, multiple-aperture photography allows post-capture manipulation of all three camera controls—aperture, shutter speed, and focus—from the same number of images used in basic HDR photography.

## 4.2 Photography by Varying Aperture

Suppose we have a set of photographs of a scene taken from the same viewpoint with different apertures, holding all other camera settings fixed. Under this scenario, image formation can be expressed in terms of four components: a scene-independent lens attenuation factor $\mathbf{R}$, a scene radiance term $\overline{\mathbf{L}}$, the sensor response function $g(\cdot)$, and image noise $\eta$,

$$\mathbf{I}(x, y, a) = g\Big( \overbrace{\underbrace{\mathbf{R}(x, y, a, f)}_{\text{lens term}} \cdot \underbrace{\overline{\mathbf{L}}(x, y, a, f)}_{\text{scene radiance term}}}^{\text{sensor irradiance}} \Big) + \underbrace{\eta}_{\text{noise}} , \tag{4.1}$$

where $\mathbf{I}(x, y, a)$ is image intensity at pixel $(x, y)$ when the aperture is $a$. In this expression, the lens term $\mathbf{R}$ models the radiometric effects of the lens and depends on pixel position, aperture, and the focus setting, $f$, of the lens. The radiance term $\overline{\mathbf{L}}$ corresponds to the mean scene radiance integrated over the aperture, *i.e.*, the total radiance subtended by aperture $a$ divided by the solid angle. We use mean radiance because this allows us to decouple the effects of exposure, which depends on aperture but is scene-independent, and of defocus, which also depends on aperture.

Given the set of captured images, our goal is to perform two operations:

- **High dynamic range photography.** Convert each of the input photos to HDR, *i.e.*, recover $\overline{\mathbf{L}}(x, y, a, f)$ for the input camera settings, $(a, f)$.
- **Post-capture aperture and focus control.** Compute $\overline{\mathbf{L}}(x, y, a', f')$ for any aperture and focus setting, $(a', f')$.

While HDR photography is straightforward by controlling exposure time rather than aperture [78], in our input photos, defocus and exposure are deeply interrelated according to the aperture setting. Hence, existing HDR and defocus analysis methods do not apply, and an entirely new inverse problem must be formulated and solved.

To do this, we establish a computationally tractable model for the terms in Eq. (4.1) that well approximates the image formation in consumer SLR digital cameras. Importantly, we show that this model leads to a restoration-based optimization problem that can be solved efficiently.

## 4.3 Image Formation Model

**Sensor model.** Following the high dynamic range literature [78], we express the sensor response $g(\cdot)$ in Eq. (4.1) as a smooth, monotonic function mapping the sensor irradiance $\mathbf{R}\cdot\overline{\mathbf{L}}$ to image intensity in the range $[0,1]$. The effective dynamic range is limited by over-saturation, quantization, and the sensor noise $\eta$, which we model as additive. Note that in Chapter 6 we consider more general models of noise.

**Exposure model.** Since we hold exposure time constant, a key factor in determining the magnitude of sensor irradiance is the size of the aperture. In particular, we represent the total solid angle subtended by the aperture with an exposure factor $e_a$, which converts between the mean radiance $\overline{\mathbf{L}}$ and the total radiance integrated over the aperture, $e_a\overline{\mathbf{L}}$. Because this factor is scene-independent, we incorporate it in the lens term,

$$\mathbf{R}(x,y,a,f) = e_a\,\hat{\mathbf{R}}(x,y,a,f)\ , \tag{4.2}$$

therefore the factor $\hat{\mathbf{R}}(x,y,a,f)$ models residual radiometric distortions, such as vignetting, that vary spatially and depend on aperture and focus setting. To resolve the multiplicative ambiguity, we assume that $\hat{\mathbf{R}}$ is normalized so the center pixel is assigned a factor of one.

**Defocus model.** While more general models are possible [11], we assume that the defocus induced by the aperture obeys the standard thin lens model [16, 92]. This model has the attractive feature that for a fronto-parallel scene, relative changes in defocus due to aperture setting are independent of depth.

In particular, for a fronto-parallel scene with radiance $\mathbf{L}$, the defocus from a given aperture can be expressed by the convolution $\overline{\mathbf{L}} = \mathbf{L} * B_\sigma$ [92]. The 2D point-spread function $B$ is parameterized by the effective *blur diameter*, $\sigma$, which depends on scene depth, focus setting, and

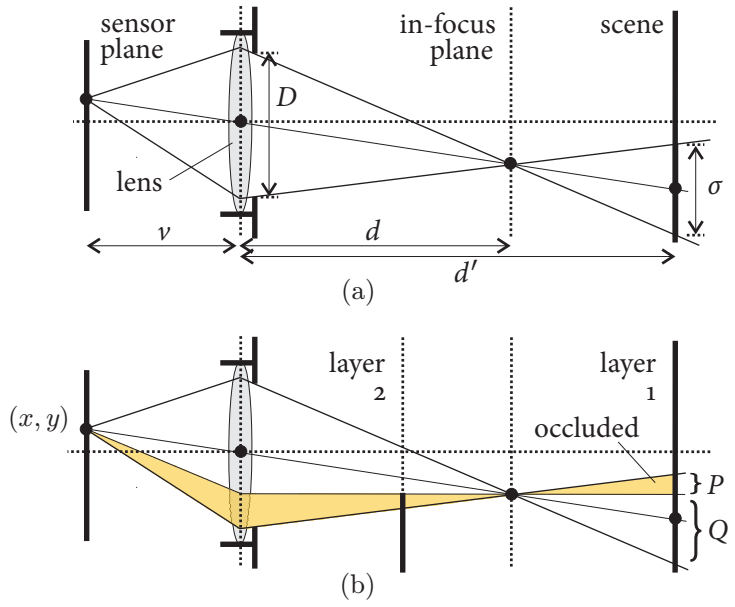



**Figure 4.2:** Defocused image formation with the thin lens model. (a) Fronto-parallel scene. (b) For a two-layered scene, the shaded fraction of the cone integrates radiance from layer 2 only, while the unshaded fraction integrates the unoccluded part of layer 1. Our occlusion model of Sec. 4.4 approximates layer 1's contribution to the radiance at $(x, y)$ as $(\mathbf{L}_P + \mathbf{L}_Q)\frac{|Q|}{|P|+|Q|}$, where $\mathbf{L}_P$ and $\mathbf{L}_Q$ represent the total radiance from regions $P$ and $Q$ respectively. This is a good approximation when $\frac{1}{|P|}\mathbf{L}_P \approx \frac{1}{|Q|}\mathbf{L}_Q$.

aperture size (Fig. 4.2a). From simple geometry,

$$\sigma = \frac{|d' - d|}{d} D \ , \tag{4.3}$$

where $d'$ is the depth of the scene, $d$ is the depth of the in-focus plane, and $D$ is the effective diameter of the aperture. This implies that regardless of the scene depth, for a fixed focus setting, the blur diameter is proportional to the aperture diameter.[1]

The thin lens geometry also implies that whatever its form, the point-spread function $B$ will scale radially with blur diameter, *i.e.*, $B_\sigma(x, y) = \frac{1}{\sigma^2} B(\frac{x}{\sigma}, \frac{y}{\sigma})$. In practice, we assume that $B_\sigma$ is a 2D symmetric Gaussian, where $\sigma$ represents the standard deviation (Sec. 2.3.5).

[1]Because it is based on simple convolution, the thin lens model for defocus implicitly assumes that scene radiance $\mathbf{L}$ is constant over the cone subtended by the largest aperture. The model also implies that any camera settings yielding the same blur diameter $\sigma$ will produce the same defocused image, *i.e.*, that generalized confocal constancy (Sec. 3.7.2) is satisfied [53].

## 4.4 Layered Scene Radiance

To make the reconstruction problem tractable, we rely on a simplified scene model that consists of multiple, possibly overlapping, fronto-parallel layers, ideally corresponding to a gross object-level segmentation of the 3D scene.

In this model, the scene is composed of $K$ layers, numbered from back to front. Each layer is specified by an HDR image, $\mathbf{L}_k$, that describes its outgoing radiance at each point, and an alpha matte, $\mathbf{A}_k$, that describes its spatial extent and transparency.

**Approximate layered occlusion model.** Although the relationship between defocus and aperture setting is particularly simple for a single-layer scene, the multiple layer case is significantly more challenging due to occlusion.[2] A fully accurate simulation of the thin lens model under occlusion involves backprojecting a cone into the scene, and integrating the unoccluded radiance (Fig. 4.2b) using a form of ray-tracing [16]. Unfortunately, this process is computationally intensive, since the point-spread function can vary with arbitrary complexity according to the geometry of the occlusion boundaries.

For computational efficiency, we therefore formulate an approximate model for layered image formation (Fig. 4.3) that accounts for occlusion, is effective in practice, and leads to simple analytic gradients used for optimization.

The model entails defocusing each scene layer independently, according to its depth, and combining the results using image compositing:

$$\overline{\mathbf{L}} = \sum_{k=1}^{K} \left[ (\mathbf{A}_k \cdot \mathbf{L}_k) * B_{\sigma_k} \right] \cdot \mathbf{M}_k \quad , \tag{4.4}$$

where $\sigma_k$ is the blur diameter for layer $k$, $\mathbf{M}_k$ is a second alpha matte for layer $k$, representing the cumulative occlusion from defocused layers in front,

$$\mathbf{M}_k \;=\; \prod_{j=k+1}^{K} \left(1 - \mathbf{A}_j * B_{\sigma_j}\right) \quad , \tag{4.5}$$

and $\cdot$ denotes pixel-wise multiplication. Eqs. (4.4) and (4.5) can be viewed as an application of the matting equation [104], and generalizes the method of McGuire, *et al.* [77] to arbitrary focus settings and numbers of layers.

Intuitively, rather than integrating partial cones of rays that are restricted by the geometry of

[2] Since we model the layers as thin, occlusion due to perpendicular step edges [22] can be ignored.

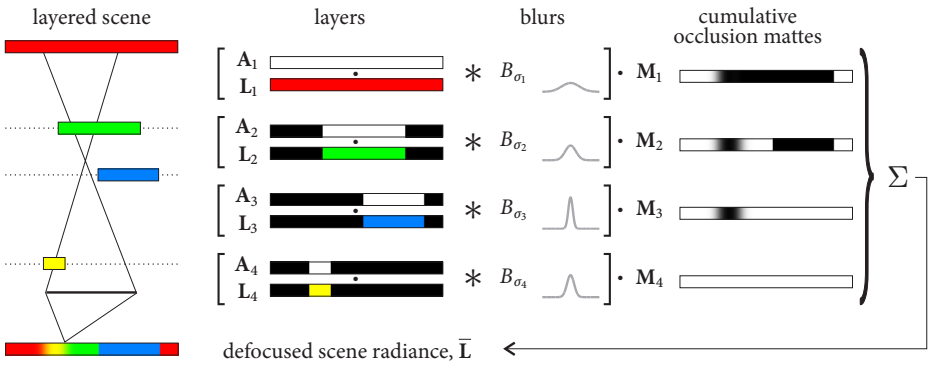



**Figure 4.3:** Approximate layered image formation model with occlusion, illustrated in 2D. The double-cone shows the thin lens geometry for a given pixel, indicating that layer 3 is nearly in-focus. To compute the defocused radiance, $\bar{\mathbf{L}}$, we use convolution to independently defocus each layer $\mathbf{A}_k \cdot \mathbf{L}_k$, where the blur diameters $\sigma_k$ are defined by the depths of the layers (Eq. (4.3)). We combine the independently defocused layers using image compositing, where the mattes $\mathbf{M}_k$ account for cumulative occlusion from defocused layers in front.

the occlusion boundaries (Fig. 4.2b), we integrate the entire cone for each layer, and weigh each layer's contribution by the fraction of rays that reach it. These weights are given by the alpha mattes, and model the thin lens geometry exactly.

In general, our approximation is accurate when the region of a layer that is subtended by the entire aperture has the same mean radiance as the unoccluded region (Fig. 4.2b). This assumption is less accurate when only a small fraction of the layer is unoccluded, but this case is mitigated by the small contribution of the layer to the overall integral. Worst-case behavior occurs when an occlusion boundary is accidentally aligned with a brightness or texture discontinuity on the occluded layer, however this is rare in practice.

**All-in-focus scene representation.** In order to simplify our formulation even further, we represent the entire scene as a single all-in-focus HDR radiance map, $\mathbf{L}$. In this reduced representation, each layer is modeled as a binary alpha matte $\mathbf{A}'_k$ that "selects" the unoccluded pixels corresponding to that layer. Note that if the narrowest-aperture input photo is all-in-focus, the brightest regions of $\mathbf{L}$ can be recovered directly, however this condition is not a requirement of our method.

While the all-in-focus radiance directly specifies the unoccluded radiance $\mathbf{A}'_k \cdot \mathbf{L}$ for each layer, to accurately model defocus near layer boundaries we must also estimate the radiance for occluded regions (Fig. 4.2b). Our underlying assumption is that $\mathbf{L}$ is sufficient to describe these occluded regions as extensions of the unoccluded layers. This allows us to apply the same image

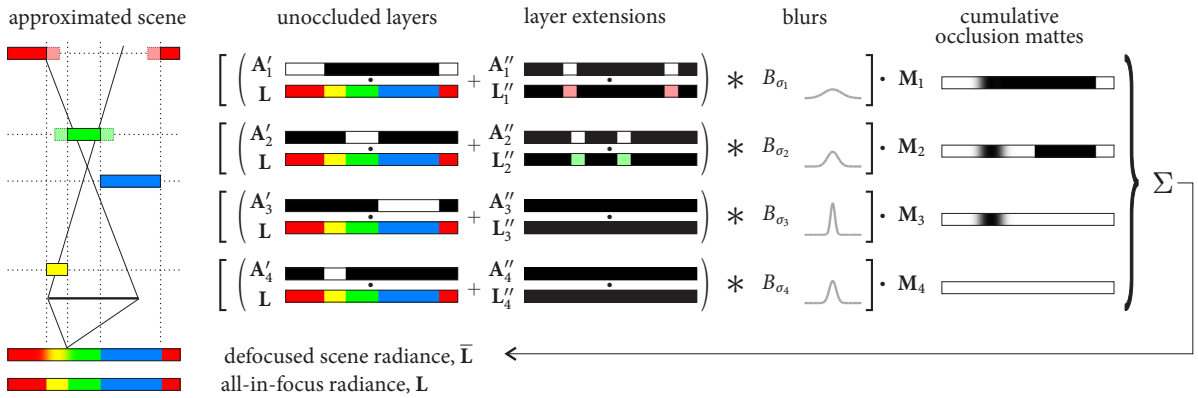



**Figure 4.4:** Reduced representation for the layered scene in Fig. 4.3, based on the all-in-focus radiance, $\mathbf{L}$. The all-in-focus radiance specifies the unoccluded regions of each layer, $\mathbf{A}'_k \cdot \mathbf{L}$, where $\{\mathbf{A}'_k\}$ is a hard segmentation of the unoccluded radiance into layers. We assume that $\mathbf{L}$ is sufficient to describe the occluded regions of the scene as well, with inpainting (lighter, dotted) used to extend the unoccluded regions behind occluders as required. Given these extended layers, $\mathbf{A}'_k \cdot \mathbf{L} + \mathbf{A}''_k \cdot \mathbf{L}''_k$, we apply the same image formation model as in Fig. 4.3.

formation model of Eqs. (4.4)–(4.5) to extended versions of the unoccluded layers (Fig. 4.4):

$$\mathbf{A}_k = \mathbf{A}'_k + \mathbf{A}''_k \tag{4.6}$$

$$\mathbf{L}_k = \mathbf{A}'_k \cdot \mathbf{L} + \mathbf{A}''_k \cdot \mathbf{L}''_k \ . \tag{4.7}$$

In Sec. 4.7 we describe our method for extending the unoccluded layers using image inpainting.

**Complete scene model.** In summary, we represent the scene by the triple $(\mathbf{L}, \mathbf{A}, \sigma)$, consisting of the all-in-focus HDR scene radiance, $\mathbf{L}$, the hard segmentation of the scene into unoccluded layers, $\mathbf{A} = \{\mathbf{A}'_k\}$, and the per-layer blur diameters, $\sigma$, specified for the widest aperture.[3]

## 4.5 Restoration-based Framework for HDR Layer Decomposition

In multiple-aperture photography we do not have any prior information about either the layer decomposition (*i.e.*, depth) or scene radiance. We therefore formulate an inverse problem whose goal is to compute $(\mathbf{L}, \mathbf{A}, \sigma)$ from a set of input photos. The resulting optimization can be viewed as a generalized image restoration problem that unifies HDR imaging and depth-from-defocus

[3] To relate the blur diameters over aperture setting, we rely on Eq. (4.3). Note that in practice we do not compute the aperture diameters directly from the f-numbers. For greater accuracy, we instead estimate the relative aperture diameters according to the calibrated exposure factors, $D_a \propto \sqrt{e_a / e_A}$.

by jointly explaining the input in terms of layered HDR radiance, exposure, and defocus.

In particular we formulate our goal as estimating $(\mathbf{L}, \mathbf{A}, \sigma)$ that best reproduces the input images, by minimizing the objective function

$$\mathcal{O}(\mathbf{L}, \mathbf{A}, \sigma) \;=\; \frac{1}{2} \sum_{a=1}^{A} \|\Delta(x, y, a)\|^2 \;+\; \lambda \, \|\mathbf{L}\|_\beta \;. \tag{4.8}$$

In this optimization, $\Delta(x, y, a)$ is the residual pixel-wise error between each input image $\mathbf{I}(x, y, a)$ and the corresponding synthesized image; $\|\mathbf{L}\|_\beta$ is a regularization term that favors piecewise smooth scene radiance; and $\lambda > 0$ controls the balance between squared image error and the regularization term.

The following equation shows the complete expression for the residual $\Delta(x, y, a)$, parsed into simpler components:

$$\Delta(x, y, a) \;=\; \min \left\{ \underbrace{e_a}_{\substack{\text{exposure} \\ \text{factor}}} \cdot \underbrace{\left[ \sum_{k=1}^{K} \left[ (\mathbf{A}'_k \cdot \mathbf{L} + \mathbf{A}''_k \cdot \mathbf{L}''_k) * B_{\sigma_{a,k}} \right] \cdot \mathbf{M}_k \right]}_{\substack{\text{layered occlusion model, } \bar{\mathbf{L}} \\ \text{from Eqs. (4.4)–(4.5)}}}, \underbrace{1}_{\substack{\text{clipping} \\ \text{term}}} \right\} \;-$$

$$\underbrace{\frac{1}{\hat{\mathbf{R}}(x, y, a, f)} \, g^{-1}\big(\mathbf{I}(x, y, a)\big)}_{\substack{\text{linearized and lens–corrected} \\ \text{image intensity}}} \;, \tag{4.9}$$

The residual is defined in terms of input images that have been linearized and lens-corrected according to pre-calibration (Sec. 4.7). This transformation simplifies the optimization of Eq. (4.8), and converts the image formation model of Eq. (4.1) to scaling by an exposure factor $e_a$, followed by clipping to model over-saturation. The innermost component of Eq. (4.9) is the layered image formation model described in Sec. 4.4.

While scaling due to the exposure factor greatly affects the relative magnitude of the additive noise, $\eta$, this effect is handled implicitly by the restoration. Note, however, that additive noise from Eq. (4.1) is modulated by the linearizing transformation that we apply to the input images, yielding modified additive noise at every pixel:

$$\eta'(x, y, a) \;=\; \frac{1}{\hat{\mathbf{R}}(x, y, a, f)} \left| \frac{\mathrm{d} g^{-1}(\mathbf{I}(x, y))}{\mathrm{d}\mathbf{I}(x, y)} \right| \eta \;, \tag{4.10}$$

where $\eta' \to \infty$ for over-saturated pixels [101].

**Weighted TV regularization.** To regularize Eq. (4.8), we use a form of the total variation (TV) norm, $\|\mathbf{L}\|_{\text{TV}} = \int \|\nabla \mathbf{L}\|$. This norm is useful for restoring sharp discontinuities, while suppressing noise and other high frequency detail [116]. The variant we propose,

$$\|\mathbf{L}\|_{\beta} = \int \sqrt{\left(w(\mathbf{L})\,\|\nabla \mathbf{L}\|\right)^2 + \beta}\ , \tag{4.11}$$

includes a perturbation term $\beta > 0$ that remains constant[4] and ensures differentiability as $\nabla \mathbf{L} \to 0$ [116]. More importantly, our norm incorporates per-pixel weights $w(\mathbf{L})$ meant to equalize the TV penalty over the high dynamic range of scene radiance (Fig. 4.12).

We define the weight $w(\mathbf{L})$ for each pixel according to its inverse exposure level, $1/e_{a^*}$, where $a^*$ corresponds to the aperture for which the pixel is "best exposed". In particular, we synthesize the transformed input images using the current scene estimate, and for each pixel we select the aperture with highest signal-to-noise ratio, computed with the noise level $\eta'$ predicted by Eq. (4.10).

## 4.6 Optimization Method

To optimize Eq. (4.8), we use a series of alternating minimizations, each of which estimates one of $\mathbf{L}$, $\mathbf{A}$, $\sigma$ while holding the rest constant.

- **Image restoration** To recover the scene radiance $\mathbf{L}$ that minimizes the objective, we take a direct iterative approach [107, 116], by carrying out a set of conjugate gradient steps. Our formulation ensures that the required gradients have straightforward analytic formulas (Appendix C).
- **Blur refinement** We use the same approach, of taking conjugate gradient steps, to optimize the blur diameters $\sigma$. Again, the required gradients have simple analytic formulas (Appendix C).
- **Layer refinement** The layer decomposition $\mathbf{A}$ is more challenging to optimize because it involves a discrete labeling, but efficient optimization methods such as graph cuts [24] are not applicable. We use a naïve approach that simultaneously modifies the layer assignment of all pixels whose residual error is more than five times the median, until convergence. Each iteration in this stage evaluates whether a change in the pixels' layer assignment leads to a reduction in the objective.
- **Layer ordering** Recall that the indexing for $\mathbf{A}$ specifies the depth ordering of the layers, from back to front. To test modifications to this ordering, we note that each blur diameter corresponds to two possible depths, either in front of or behind the in-focus plane

[4] We used $\beta = 10^{-8}$ in all our experiments.

(Eq. (4.3), Sec. 2.7). We use a brute force approach that tests all $2^{K-1}$ distinct layer orderings, and select the one leading to the lowest objective (Fig. 4.6d).

Note that even when the layer ordering and blur diameters are specified, a two-fold ambiguity still remains. In particular, our defocus model alone does not let us resolve whether the layer with the smallest blur diameter (*i.e.*, the most in-focus layer) is in front of or behind the in-focus plane. In terms of resynthesizing new images, this ambiguity has little impact provided that the layer with the smallest blur diameter is nearly in focus. For greater levels of defocus, however, the ambiguity can be significant. Our current approach is to break the ambiguity arbitrarily, but we could potentially analyze errors at occlusion boundaries or exploit additional information (*e.g.*, that the lens is focused behind the scene [111]) to resolve this.

- **Initialization** In order for this procedure to work, we need to initialize all three of $(\mathbf{L}, \mathbf{A}, \sigma)$ with reasonable estimates, as discussed below.

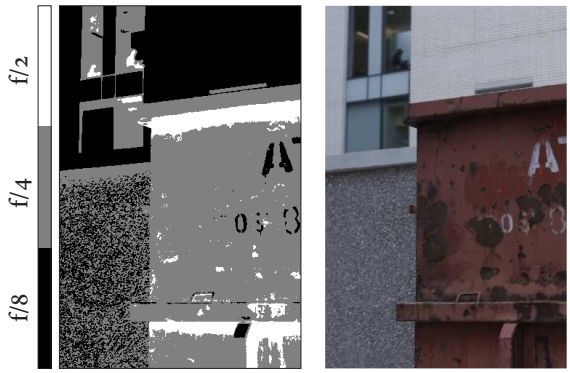
## 4.7 Implementation Details

**Scene radiance initialization.** We define an initial estimate for the unoccluded radiance, **L**, by directly selecting pixels from the transformed input images, then scaling them by their inverse exposure factor, $1/e_a$, to convert them to HDR radiance. Our strategy is to select as many pixels as possible from the sharply focused narrowest-aperture image, but to make adjustments for darker regions of the scene, whose narrow-aperture image intensities will be dominated by noise (Fig. 4.5).

For each pixel, we select the narrowest aperture for which the image intensity is above a fixed threshold of $\kappa = 0.1$, or if none meet this threshold, then we select the largest aperture. In terms of Eq. (4.10), the threshold defines a minimum acceptable signal-to-noise ratio of $\kappa/\eta'$.

**Initial layering and blur assignment.** To obtain an initial estimate for the layers and blur diameters, we use a simple window-based depth-from-defocus method inspired by classic approaches [29, 92] and more recent MRF-based techniques [10, 95]. Our method involves directly testing a set of hypotheses for blur diameter, $\{\hat{\sigma}_i\}$, by synthetically defocusing the image as if the whole scene were a single fronto-parallel surface. We specify these hypotheses for blur diameter in the widest aperture, recalling that Eq. (4.3) relates each such hypothesis over all aperture settings.

Because of the large exposure differences between photos taken several f-stops apart, we restrict our evaluation of consistency with a given blur hypothesis, $\hat{\sigma}_i$, to adjacent pairs of images captured with successive aperture settings, $(a, a+1)$.




source aperture, initial radiance (a)

initial radiance (tone-mapped HDR) (b)

**Figure 4.5:** Initial estimate for unoccluded scene radiance. (a) Source aperture from the input sequence, corresponding to the narrowest aperture with acceptable SNR. (b) Initial estimate for HDR scene radiance, shown using tone-mapping.

To evaluate consistency for each such pair, we use the hypothesis to align the narrower aperture image to the wider one, then directly measure per-pixel resynthesis error. This alignment involves convolving the narrower aperture image with the required incremental blur, scaling the image intensity by a factor of $e_{a+1}/e_a$, and clipping any oversaturated pixels. Since our point-spread function is Gaussian, this incremental blur can be expressed in a particularly simple form, namely another 2D symmetric Gaussian with a standard deviation of $\sqrt{{D_{a+1}}^2 - {D_a}^2}\ \hat{\sigma}_i$.

By summing the resynthesis error across all adjacent pairs of apertures, we obtain a rough per-pixel metric describing consistency with the input images over our set of blur diameter hypotheses. While this error metric can be minimized in a greedy fashion for every pixel (Fig. 4.6a), we a use Markov random field (MRF) framework to reward piecewise smoothness and recover a small number of layers (Fig. 4.6b). In particular, we employ graph cuts with the expansion-move approach [25], where the smoothness cost is defined as a truncated linear function of adjacent label differences on the four-connected grid,

$$\sum_{(x',y') \in \mathrm{neigh}(x,y)} \max\left( |l(x',y') - l(x,y)|,\ s_{\max} \right) \quad , \tag{4.12}$$

where $l(x, y)$ represents the discrete index of the blur hypothesis $\hat{\sigma}_i$ assigned to pixel $(x, y)$, and $\mathrm{neigh}(x, y)$ defines the adjacency structure. In all our experiments we used $s_{\max} = 2$.

After finding the MRF solution, we apply simple morphological post-processing to detect pixels belonging to very small regions, constituting less than 5 % of the image area, and to relabel them according to their nearest neighboring region above this size threshold. Note that our

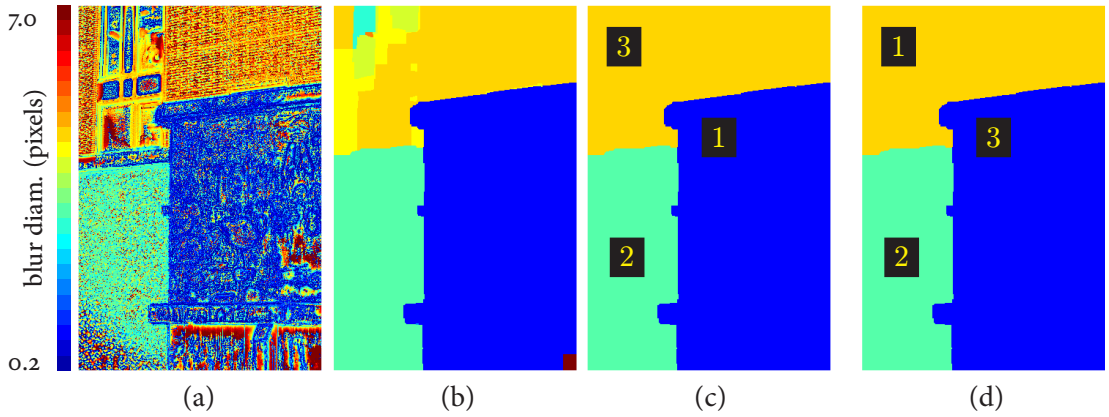


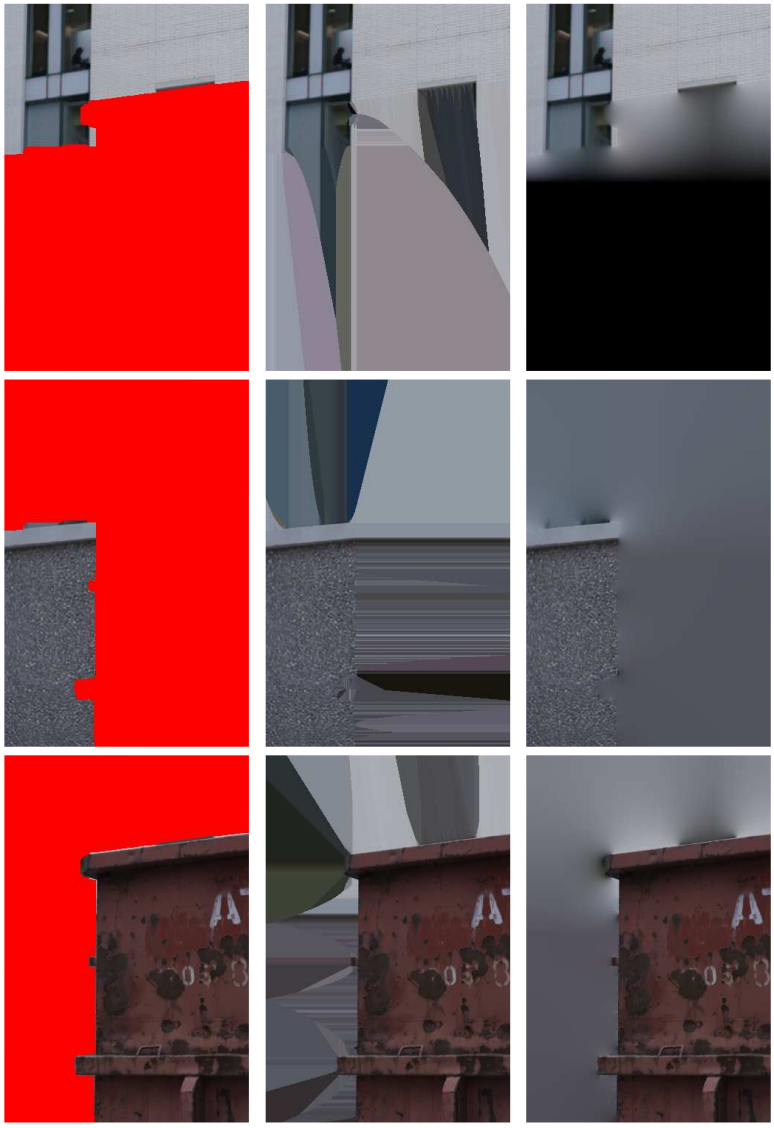

**Figure 4.6:** (a)–(c) Initial layer decomposition and blur assignment for the DUMPSTER dataset, computed using our depth-from-defocus method. (a) Greedy layer assignment. (b) MRF-based layer assignment. (c) Initial layer decomposition, determined by applying morphological post-processing to (b). Our initial guess for the back-to-front depth ordering is also shown. (d) Final layering, which involves re-estimating the depth ordering and iteratively modifying the layer assignment for high-residual pixels. The corrected depth ordering significantly improves the quality of resynthesis, however the effect of modifying the layer assignment is very subtle.

implementation currently assumes that all pixels assigned to the same blur hypothesis belong the same depth layer. While this simplifying assumption is appropriate for all our examples (*e.g.*, the two window panes in Fig. 4.14) and limits the number of layers, a more general approach is to assign disconnected regions of pixels to separate layers (we did not do this in our implementation).

**Sensor response and lens term calibration.** To recover the sensor response function, $g(\cdot)$, we apply standard HDR imaging methods [78] to a calibration sequence captured with varying exposure time.

We recover the radiometric lens term $\mathbf{R}(x, y, a, f)$ using one-time pre-calibration process as well. To do this, we capture a calibration sequence of a diffuse and textureless plane, and take the pixel-wise approach described in Sec. 3.5. In practice our implementation ignores the dependence of $\mathbf{R}$ on focus setting, but if the focus setting is recorded at capture time, we can use it to interpolate over a more detailed radiometric calibration measured over a range of focus settings (Sec. 3.5).

**Occluded radiance estimation.** As illustrated in Fig. 4.4, we assume that all scene layers can be expressed in terms of the unoccluded all-in-focus radiance $\mathbf{L}$. During optimization, we use a simple inpainting method to extend the unoccluded layers: we use a naïve, low-cost tech-



masked out background (a) | inpainted background (nearest pixel) (b) | inpainted background (diffusion) (c)

**Figure 4.7:** Layering and background inpainting for the Dumpster dataset. (a) The three recovered scene layers, visualized by masking out the background. (b) Inpainting the background for each layer using the nearest layer pixel. (c) Using diffusion-based inpainting [21] to define the layer background. In practice, we need not compute the inpainting for the front-most layer (bottom row).

nique that extends each layer by filling its occluded background with the closest unoccluded pixel from its boundary (Fig. 4.7b). For synthesis, however, we obtain higher-quality results by using a simple variant of PDE-based inpainting [21] (Fig. 4.7c), which formulates inpainting as a diffusion process. Previous approaches have used similar inpainting methods for synthesis [69, 77], and have also explored using texture synthesis to extend the unoccluded layers [79].

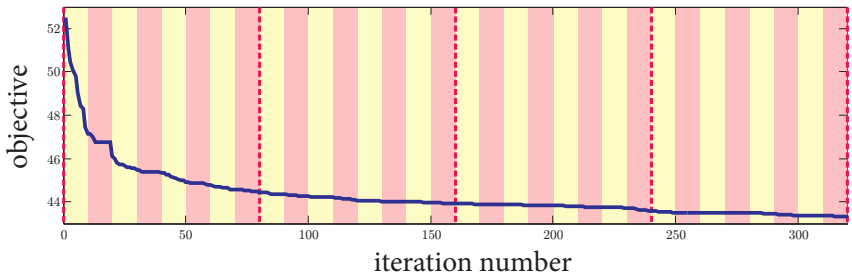



**Figure 4.8:** Typical convergence behavior of our restoration method, shown for the DUMPSTER dataset (Fig. 4.1). The yellow and pink shaded regions correspond to alternating blocks of image restoration and blur refinement respectively (10 iterations each), and the dashed red vertical lines indicate layer reordering and refinement (every 80 iterations).

## 4.8 Results and Discussion

To evaluate our approach we captured several real datasets using two different digital SLR cameras. We also generated a synthetic dataset to enable comparison with ground truth (LENA dataset).

We captured the real datasets using the Canon EOS-1Ds Mark II (DUMPSTER, PORTRAIT, MACRO datasets) or the EOS-1Ds Mark III (DOORS dataset), secured on a sturdy tripod. In both cases we used a wide-aperture fixed focal length lens, the Canon EF85mm f1.2L and the EF50mm f1.2L respectively, set to manual focus. For all our experiments we used the built-in three-image "aperture bracketing" mode set to ±2 stops, and chose the shutter speed so that the images were captured at f/8, f/4, and f/2 (yielding relative exposure levels of roughly 1, 4, and 16). We captured 14-bit RAW images for increased dynamic range, and demonstrate our method for downsampled images with resolutions of $500 \times 333$ or $705 \times 469$ pixels.[5]

Our image restoration algorithm follows the description in Sec. 4.6, alternating between 10 conjugate gradient steps each of image restoration and blur refinement, until convergence. We periodically apply the layer reordering and refinement procedure as well, both immediately after initialization and every 80 such steps. As Fig. 4.8 shows, the image restoration typically converges within the first 100 iterations, and beyond the first application, layer reordering and refinement has little effect. For all experiments we set the smoothing parameter to $\lambda = 0.002$.

**Resynthesis with new camera settings.** Once the image restoration has been computed, *i.e.*, once $(\mathbf{L}, \mathbf{A}, \sigma)$ has been estimated, we can apply the forward image formation model with arbitrary camera settings, and resynthesize new images at near-interactive rates (Figs. 4.1,4.9–

[5]For additional results and videos, see `http://www.cs.toronto.edu/~hasinoff/aperture/`.

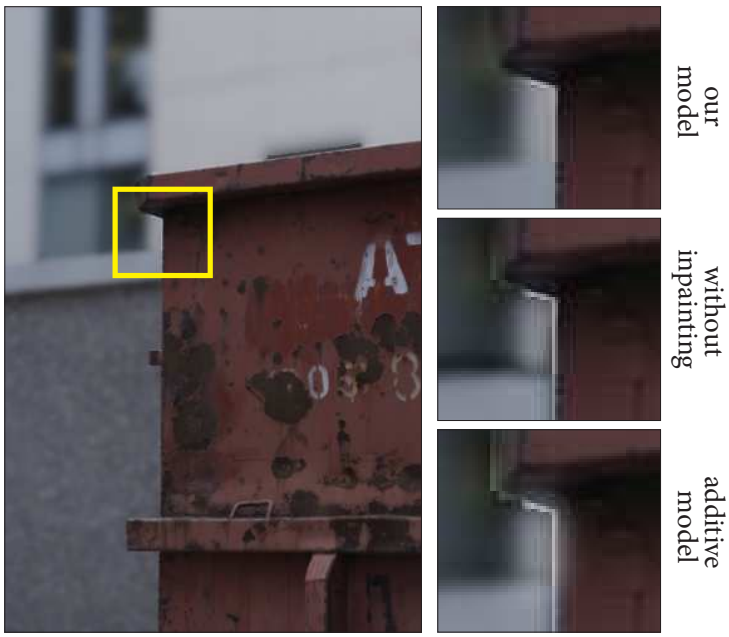



**Figure 4.9:** Layered image formation results at occlusion boundaries. *Left:* Tone-mapped HDR image of the Dumpster dataset, for an extrapolated aperture (f/1). *Top inset:* Our model handles occlusions in a visually realistic way. *Middle:* Without inpainting, *i.e.*, assuming zero radiance in occluded regions, the resulting darkening emphasizes pixels whose layer assignment has been misestimated, that are not otherwise noticeable. *Bottom:* An additive image formation model [95, 107] exhibits similar artifacts, plus erroneous spill from the occluded background layer.

4.17).[6] Note that since we do not record the focus setting $f$ at capture time, we fix the in-focus depth arbitrarily (*e.g.*, to 1.0 m), which allows us to specify the layer depths in relative terms (Fig. 4.17). To synthesize photos with modified focus settings, we express the depth of the new focus setting as a fraction of the in-focus depth.[7]

Note that while camera settings can also be extrapolated, this functionality is somewhat limited. In particular, while extrapolating larger apertures than lets us model exposure changes and increased defocus for each depth layer (Fig. 4.9), the depth resolution of our layered model is limited compared to what larger apertures could potentially provide [99].

To demonstrate the benefit of our layered occlusion model for resynthesis, we compared our resynthesis results at layer boundaries with those obtained using alternative methods. As shown in Fig. 4.9, our layered occlusion model produces visually realistic output, even in the absence of pixel-accurate layer assignment. Our model is a significant improvement over the

[6] In order to visualize the exposure range of the recovered HDR radiance, we apply tone-mapping using a simple global operator of the form $T(x) = \frac{x}{1+x}$.

[7] For ease of comparison, when changing the focus setting synthetically, we do not resynthesize geometric distortions such as image magnification (Sec. 3.6). Similarly, we do not simulate the residual radiometric distortions $\hat{\mathbf{R}}$, such as vignetting. All these lens-specific artifacts can be simulated if desired.


**Figure 4.10:** Synthetic LENA dataset. *Left:* Underlying 3D scene model, created from an HDR version of the Lena image. *Right:* Input images generated by applying our image formation model to the known 3D model, focused on the middle layer.

typical additive model of defocus [95, 107], which shows objectionable rendering artifacts at layer boundaries. Importantly, our layered occlusion model is accurate enough that we can resolve the correct layer ordering in all our experiments (except for one error in the DOORS dataset), simply by applying brute force search and testing which ordering leads to the smallest objective.

**Synthetic data: LENA dataset.** To enable comparison with ground truth, we tested our approach using a synthetic dataset (Fig. 4.10). This dataset consists of an HDR version of the 512 × 512 pixel Lena image, where we simulate HDR by dividing the image into three vertical bands and artificially exposing each band. We decomposed the image into layers by assigning different depths to each of three horizontal bands, and generated the input images by applying the forward image formation model, focused on the middle layer. Finally, we added Gaussian noise to the input with a standard deviation of 1 % of the intensity range.

As Fig. 4.11 shows, the restoration and resynthesis agree well with the ground truth, and show no visually objectionable artifacts, even at layer boundaries. The results show denoising throughout the image and ev demonstrate good performance in regions that are both dark and defocused. Such regions constitute a worst case for our method, since they are dominated by noise for narrow apertures, but are strongly defocused for wide apertures. Despite the challenge presented by these regions, our image restoration framework handles them naturally, because our formulation with TV regularization encourages the "deconvolution" of blurred intensity edges while simultaneously suppressing noise (Fig. 4.12a, inset). In general, however, weaker high-frequency detail cannot be recovered from strongly defocused regions.

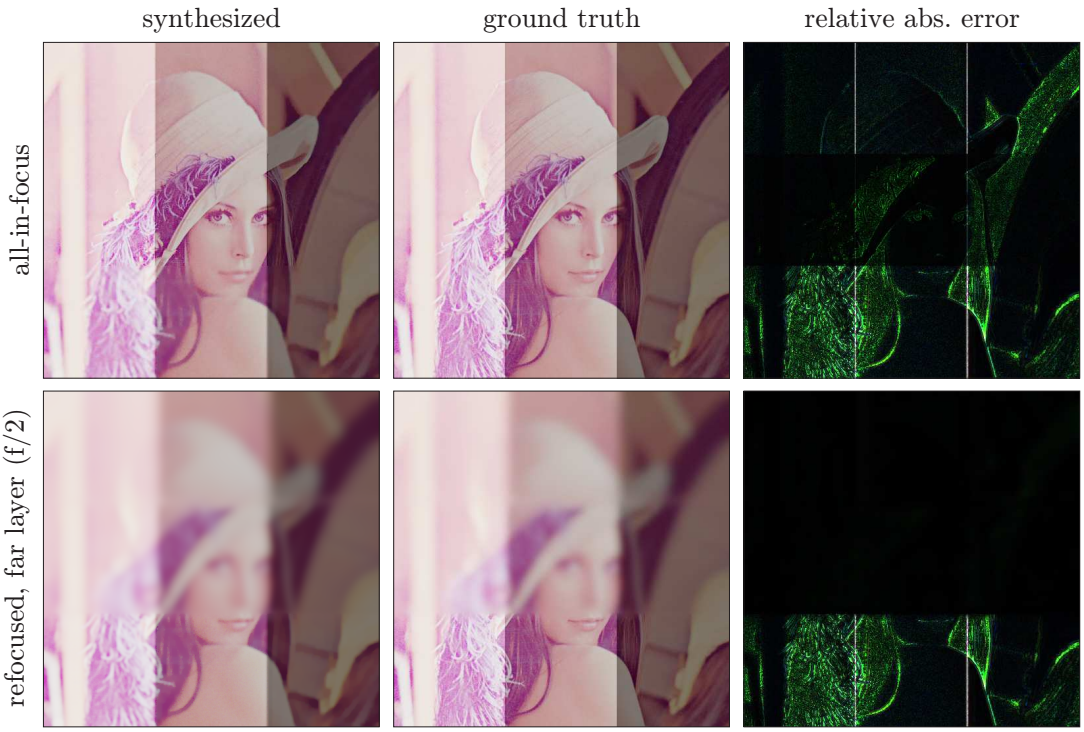



**Figure 4.11:** Resynthesis results for the Lena dataset, shown tone-mapped, agree visually with ground truth. Note the successful smoothing and sharpening . The remaining errors are mainly due to the loss of the highest frequency detail caused by our image restoration and denoising. Because of the high dynamic range, we visualize the error in relative terms, as a fraction of the ground truth radiance.


**Figure 4.12:** Effect of TV weighting. We show the all-in-focus HDR restoration result for the Lena dataset, tone-mapped and with enhanced contrast for the inset: (a) weighting the TV penalty according to effective exposure using Eq. (4.11), and (b) without weighting. In the absence of TV weighting, dark scene regions give rise to little TV penalty, and therefore get relatively under-smoothed. In both cases, TV regularization shows characteristic blocking into piecewise smooth regions.

We also used this dataset to test the effect of using different numbers of input images spanning the same range of apertures from f/8 to f/2 (Table 4.1). As Fig. 4.13 shows, using only 2 input images significantly deteriorates the restoration results. As expected, using more input images improves the restoration, particularly with respect to recovering detail in dark and defocused regions, which benefit from the noise reduction that comes from additional images.

**DUMPSTER dataset.** This outdoor scene has served as a running example throughout the chapter (Figs. 4.1, 4.5-4.9). It is composed of three distinct and roughly fronto-parallel layers: a background building, a pebbled wall, and a rusty dumpster. The foreground dumpster is darker than the rest of the scene and is almost in-focus. Although the layering recovered by the restoration is not pixel-accurate at the boundaries, resynthesis with new camera settings yields visually realistic results (Figs. 4.1 and 4.9).

**PORTRAIT dataset.** This portrait was captured indoors in a dark room, using only available light from the background window (Fig. 4.14). The subject is nearly in-focus and very dark compared to the background buildings outside, and an even darker chair sits defocused in the foreground. Note that while the final layer assignment is only roughly accurate (*e.g.*, near the subject's right shoulder), the discrepancies are restricted mainly to low-texture regions near layer boundaries, where layer membership is ambiguous and has little influence on resynthesis. In this sense, our method is similar to image-based rendering from stereo [45, 137] where reconstruction results that deviate from ground truth in "unimportant" ways can still lead to visually realistic new images. Slight artifacts can be observed at the boundary of the chair, in the form of an over-sharpened dark stripe running along its arm. This part of the scene was under-exposed even in the widest-aperture image, and the blur diameter was apparently estimated too high,

**Table 4.1:** Restoration error for the LENA dataset, using different numbers of input images spanning the aperture range f/8–f/2. All errors are measured with respect to the ground truth HDR all-in-focus radiance.

| num. input images | f-stops apart | RMS error (all-in-focus) | RMS rel. error (all-in-focus) | median rel. error (all-in-focus) |
|---|---|---|---|---|
| 2 | 4 | 0.0753 | 13.2 % | 2.88 % |
| 3 | 2 | 0.0737 | 11.7 % | 2.27 % |
| 5 | 1 | 0.0727 | 11.4 % | 1.97 % |
| 9 | 1/2 | 0.0707 | 10.8 % | 1.78 % |
| 13 | 1/3 | 0.0688 | 10.6 % | 1.84 % |

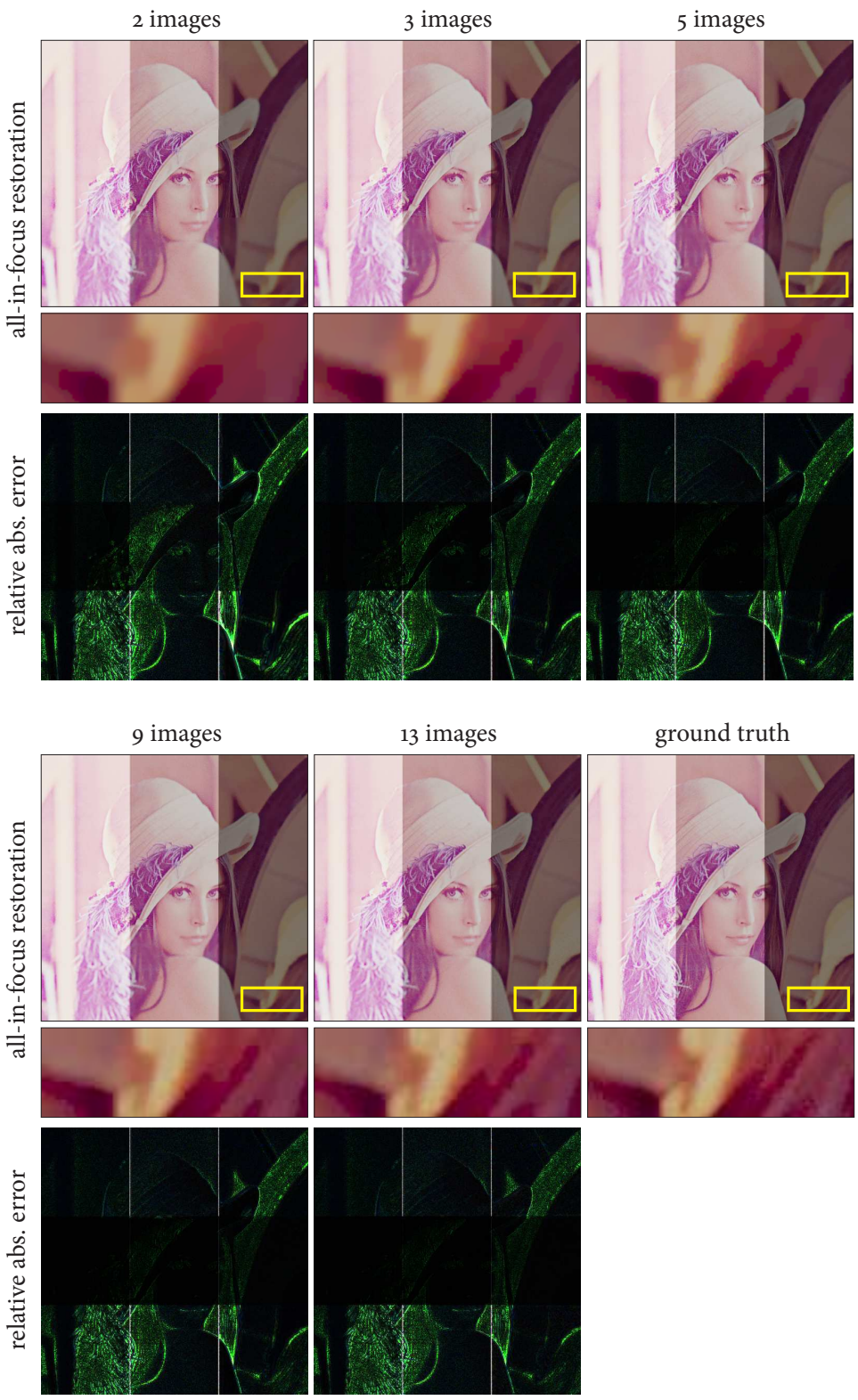



**Figure 4.13:** Effect of the number of input images for the Lena dataset. *Top of row:* Tone-mapped all-in-focus HDR restoration. For better visualization, the inset is shown with enhanced contrast. *Bottom of row:* Relative absolute error, compared to the ground truth in-focus HDR radiance.

perhaps due to over-fitting the background pixels that were incorrectly assigned to the chair.

**DOORS dataset.** This architectural scene was captured outdoors at twilight, and consists of a sloping wall containing a row of rusty doors, with a more brightly illuminated background (Fig. 4.15). The sloping, hallway-like geometry constitutes a challenging test for our method's ability to handle scenes that violate our piecewise fronto-parallel scene model. As the results show, despite the fact that our method decomposes the scene into six fronto-parallel layers, the recovered layer ordering is almost correct, and our restoration allows us to resynthesize visually realistic new images. Note that the reduced detail for the tree in the background is due to scene motion caused by wind over the 1 s total capture time.

**Failure case: MACRO dataset.** Our final sequence was a macro still life scene, captured using a 10 mm extension tube to reduce the minimum focusing distance of the lens, and to increase the magnification to approximately life-sized (1:1). The scene is composed of a miniature glass bottle whose inner surface is painted, and a dried bundle of green tea leaves (Fig. 4.16). This is a challenging dataset for several reasons: the level of defocus is severe outside the very narrow depth of field, the scene consists of both smooth and intricate geometry (bottle and tea leaves, respectively), and the reflections on the glass surface only become focused at incorrect virtual depths. The initial segmentation leads to a very coarse decomposition into layers, which is not improved by our optimization. While the resynthesis results for this scene suffer from strong artifacts, the gross structure, blur levels, and ordering of the scene layers are still recovered correctly. The worst artifacts are the bright "cracks" occurring at layer boundaries, due to a combination of incorrect layer segmentation and our diffusion-based inpainting method.

A current limitation of our method is that our scheme for re-estimating the layering is not always effective, since residual error in reproducing the input images may not be discriminative enough to identify pixels with incorrect layer labels, given overfitting and other sources of error such as imperfect calibration. Fortunately, even when the layering is not estimated exactly, our layered occlusion model often leads to visually realistic resynthesized images (*e.g.*, Figs. 4.9 and 4.14).

## Summary

We demonstrated how multiple-aperture photography leads to a unified restoration framework for decoupling the effects of defocus and exposure, which permits post-capture control of the camera settings in HDR. From a user interaction perspective, one can imagine creating new

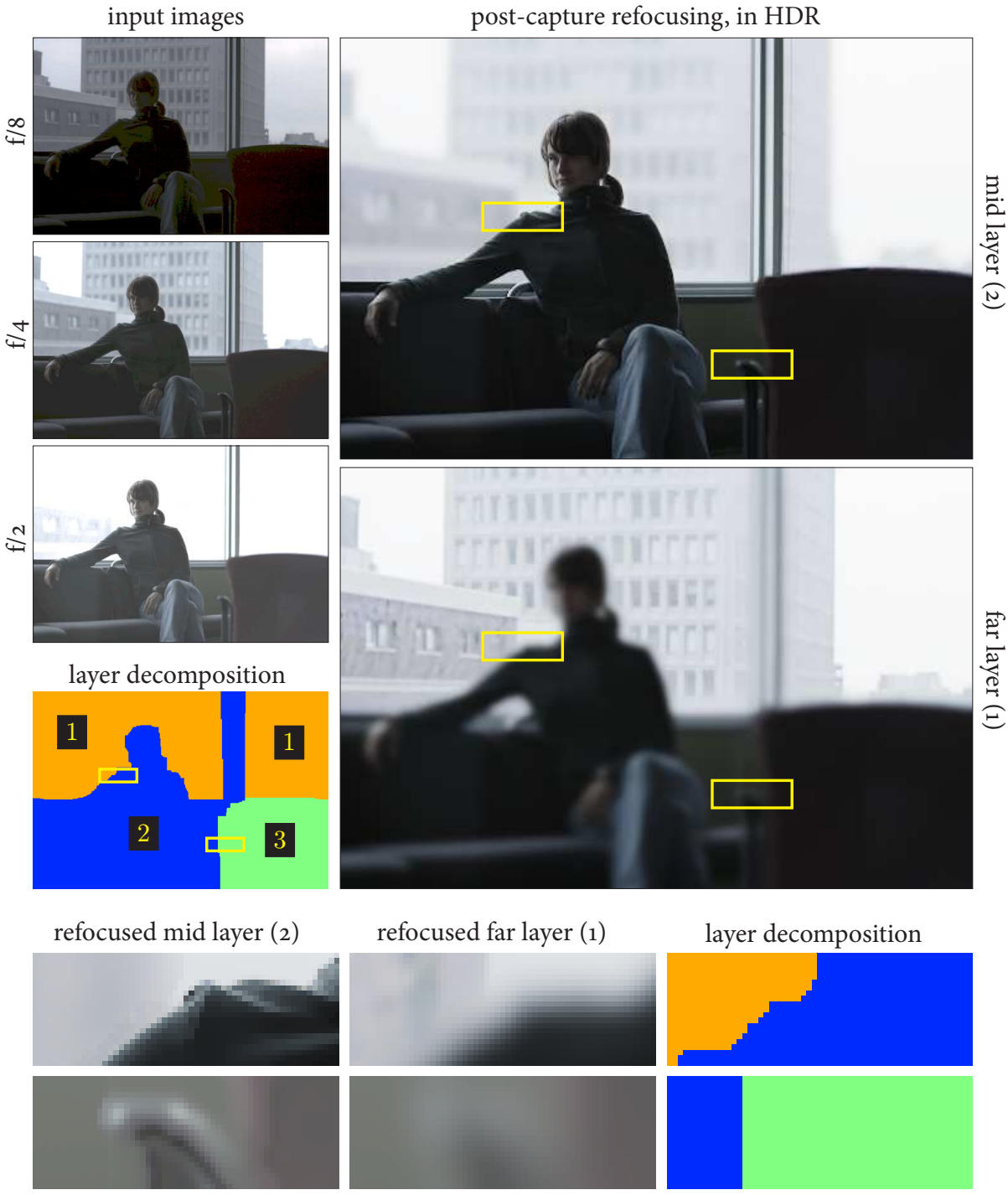



**Figure 4.14:** Portrait dataset. The input images are visualized with strong gamma correction ($\gamma = 3$) to display the high dynamic range of the scene, and show significant posterization artifacts. Although the final layer assignment has errors in low-texture regions near layer boundaries, the restoration results are sufficiently accurate to resynthesize visually realistic new images. We demonstrate refocusing in HDR with tone-mapping, simulating the widest input aperture (f/2).

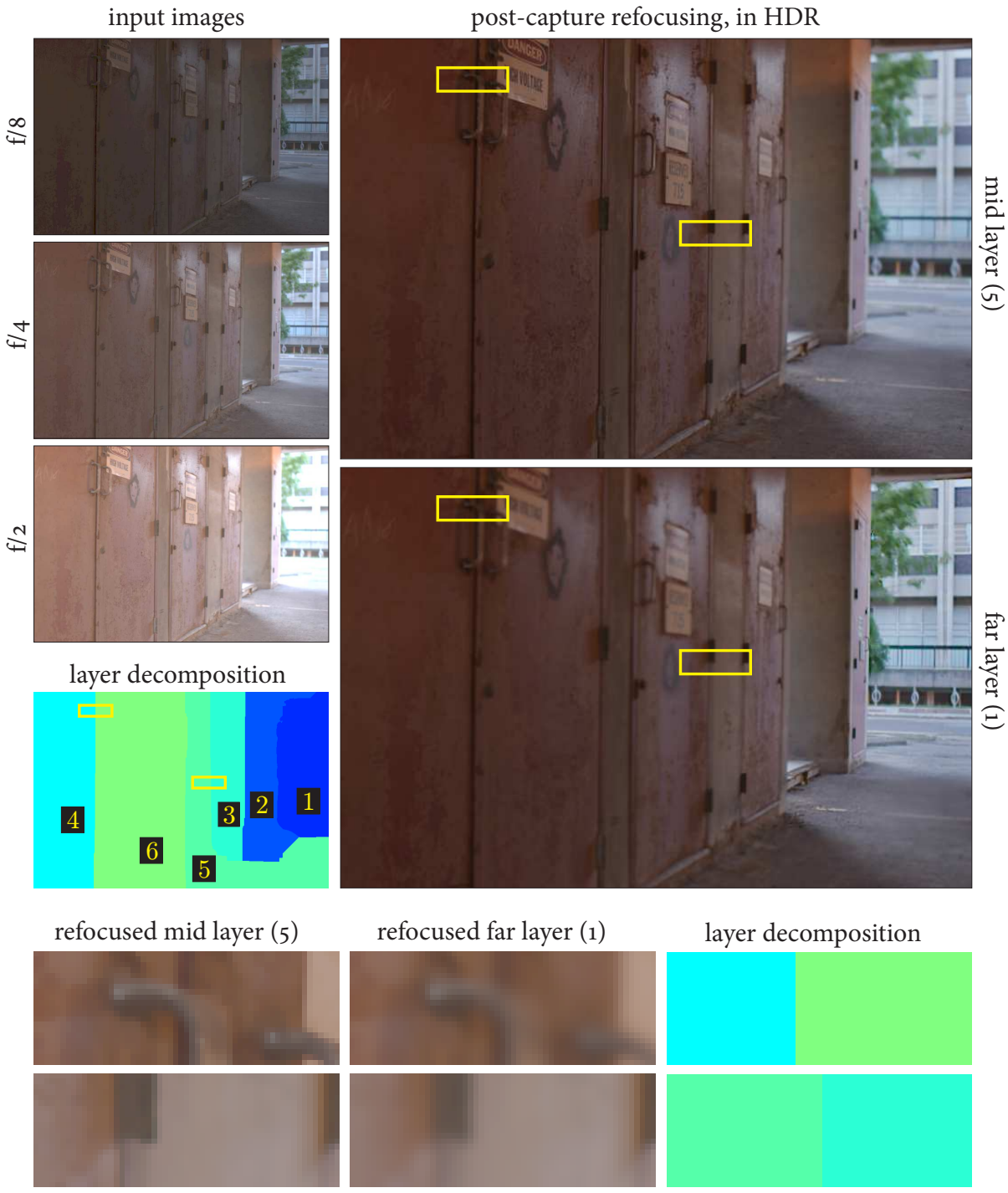



**Figure 4.15:** DOORS dataset. The input images are visualized with strong gamma correction ($\gamma = 3$) to display the high dynamic range of the scene. Our method approximates the sloping planar geometry of the scene using a small number of fronto-parallel layers. Despite this approximation, and an incorrect layer ordering estimated for the leftmost layer, our restoration results are able to resynthesize visually realistic new images. We demonstrate refocusing in HDR with tone-mapping, simulating the widest input aperture (f/2).

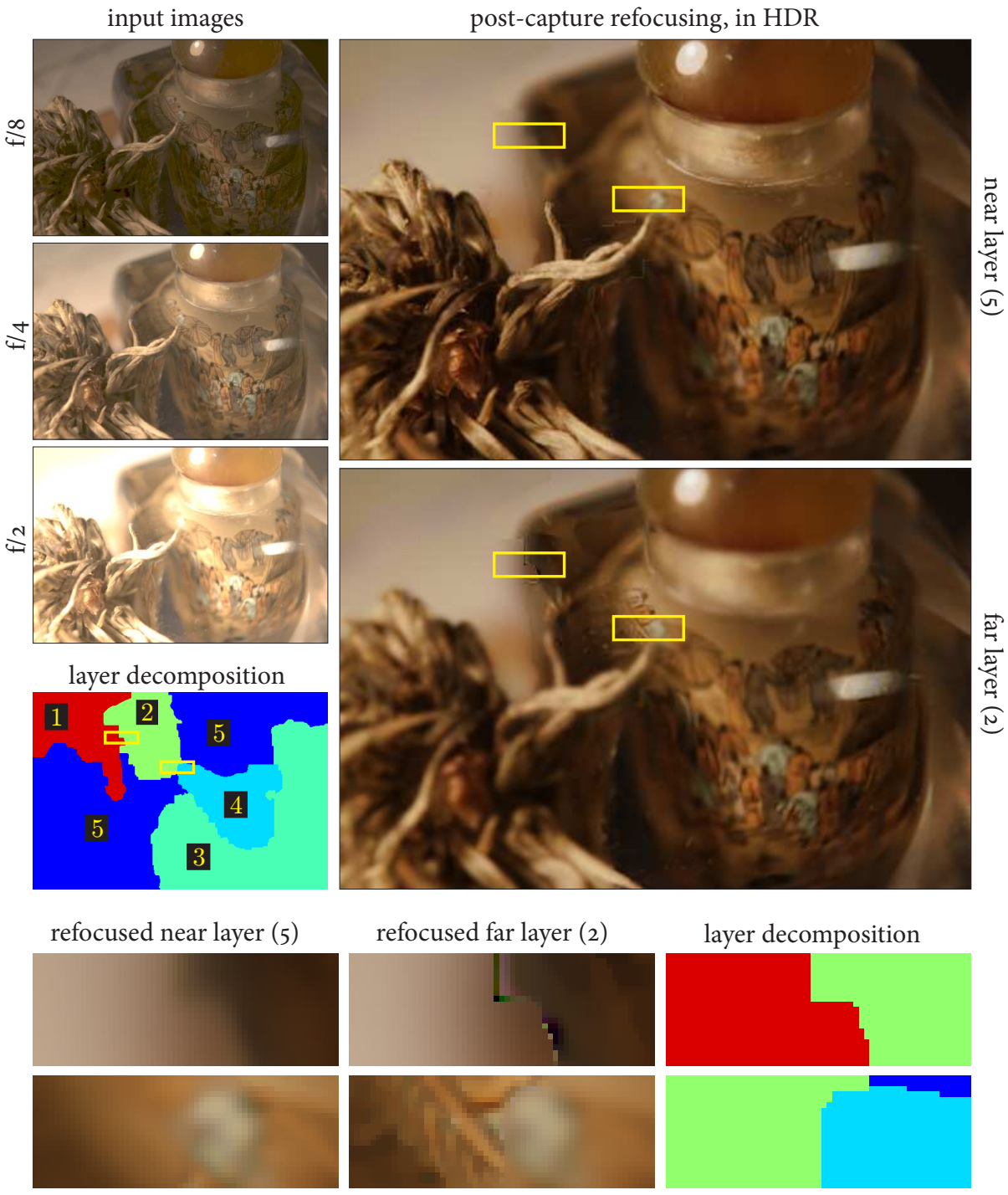



**Figure 4.16:** Macro dataset (failure case). The input images are visualized with strong gamma correction ($\gamma = 3$) to display the high dynamic range of the scene. The recovered layer segmentation is very coarse, and significant artifacts are visible at layer boundaries, due to a combination of the incorrect layer segmentation and our diffusion-based inpainting. We demonstrate refocusing in HDR with tone-mapping, simulating the widest input aperture (f/2).

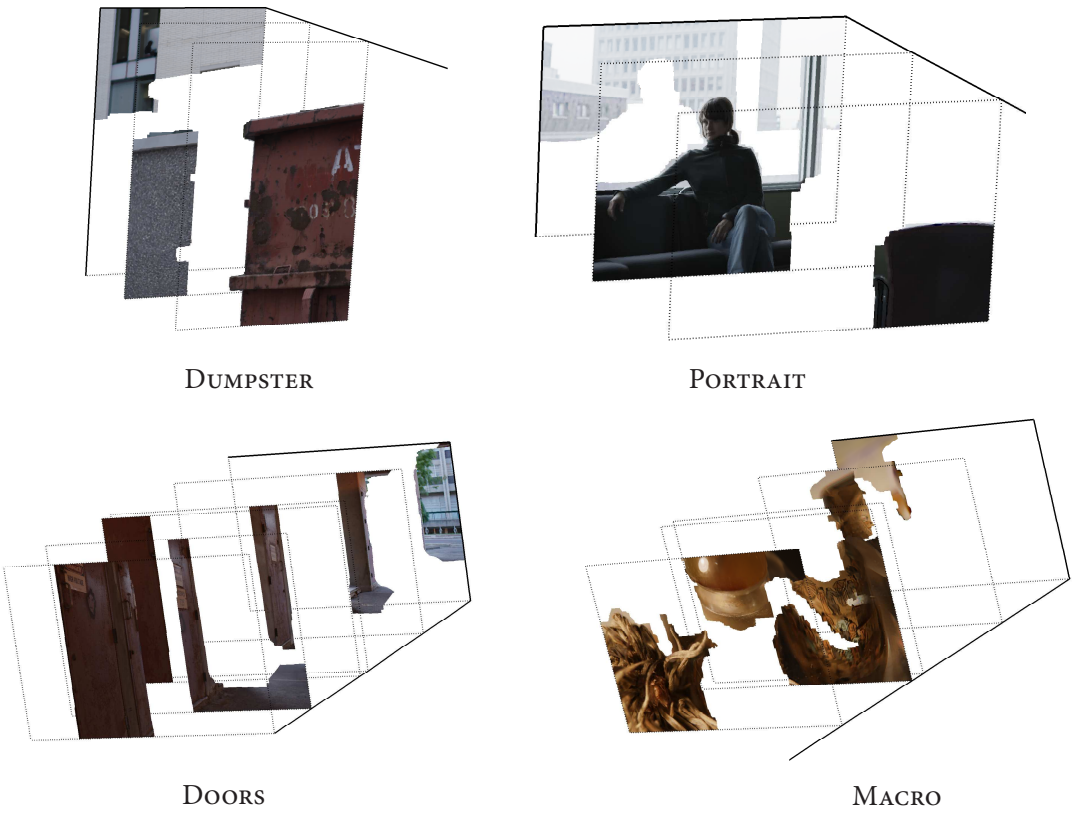



**Figure 4.17:** Gallery of restoration results for the real datasets. We visualize the recovered layers in 3D using the relative depths defined by their blur diameters and ordering.

controls to navigate the space of camera settings offered by our representation. In fact, our recovered scene model is rich enough to support non-physically based models of defocus as well, and to permit additional special effects such as compositing new objects into the scene.

For future work, we are interested in addressing motion between exposures that may be caused by hand-held photography or subject motion. Although we have experimented with simple image registration methods, it could be beneficial to integrate a layer-based parametric model of optical flow directly into the overall optimization. We are also interested in improving the efficiency of our technique by extending it to multi-resolution.

While each layer is currently modeled as a binary mask, it could be useful to represent each layer with fractional alpha values, for improved resynthesis at boundary pixels that contain mixtures of background and foreground. Our image formation model (Sec. 4.4) already handles layers with general alpha mattes, and it should be straightforward to process our layer estimates in the vicinity of the initial hard boundaries using existing matting techniques [52, 137]. This color-based matting may also be useful help refine the initial layering we estimate using depth-from-defocus.

# Chapter 5

# Light-Efficient Photography

> Efficiency is doing better what is already being done.
>
> Peter Drucker (1909–2005)

> I'll take fifty percent efficiency to get one hundred percent loyalty.
>
> Samuel Goldwyn (1879–1974)

In this chapter we consider the problem of imaging a scene with a given depth of field at a given exposure level in the shortest amount of time possible. We show that by (1) collecting a sequence of photos and (2) controlling the aperture, focus and exposure time of each photo individually, we can span the given depth of field in less total time than it takes to expose a single narrower-aperture photo. Using this as a starting point, we obtain two key results. First, for lenses with continuously-variable apertures, we derive a closed-form solution for the *globally optimal* capture sequence, *i.e.*, that collects light from the specified depth of field in the most efficient way possible. Second, for lenses with discrete apertures, we derive an integer programming problem whose solution is the optimal sequence. Our results are applicable to off-the-shelf cameras and typical photography conditions, and advocate the use of dense, wide-aperture photo sequences as a light-efficient alternative to single-shot, narrow-aperture photography.

## 5.1 Introduction

Two of the most important choices when taking a photo are the photo's exposure level and its depth of field. Ideally, these choices will result in a photo whose subject is free of noise or pixel saturation [54, 56], and appears to be in focus. These choices, however, come with a severe time constraint: in order to take a photo that has both a specific exposure level and a specific depth of field, we must expose the camera's sensor for a length of time that is dictated by the lens

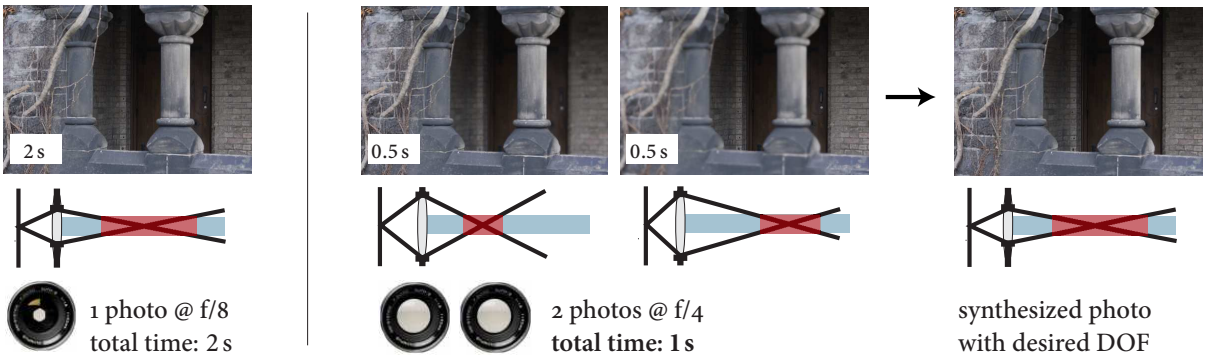



**Figure 5.1:** *Left:* Traditional single-shot photography. The desired depth of field is shown in red. *Right:* Light-efficient photography. Two wide-aperture photos span the same DOF as a single-shot narrow-aperture photo. Each wide-aperture photo requires 1/4 the time to reach the exposure level of the narrow-aperture photo, resulting in a 2× net speedup for the total exposure time.

optics. Moreover, the wider the depth of field, the longer we must wait for the sensor to reach the chosen exposure level. In practice, this makes it impossible to efficiently take sharp and well-exposed photos of a poorly-illuminated subject that spans a wide range of distances from the camera. To get a good exposure level, we must compromise something—either use a narrow depth of field (and incur defocus blur [58, 64, 92, 120]) or take a long exposure (and incur motion blur [96, 113, 131]).

In this chapter we seek to overcome the time constraint imposed by lens optics, by capturing a sequence of photos rather than just one. We show that if the aperture, exposure time, and focus setting of each photo is selected appropriately, we can span a given depth of field with a given exposure level *in less total time than it takes to expose a single photo* (Fig. 5.1). This novel observation is based on a simple fact: even though wide apertures have a narrow depth of field (DOF), they are much more efficient than narrow apertures in gathering light from within their depth of field. Hence, even though it is not possible to span a wide DOF with a single wide-aperture photo, it is possible to span it with several of them, and do so very efficiently.

Using this observation as a starting point, we develop a general theory of *light-efficient photography* that addresses four questions: (1) under what conditions is capturing photo sequences with "synthetic" DOFs more efficient than single-shot photography? (2) How can we characterize the set of sequences that are *globally optimal* for a given DOF and exposure level, *i.e.*, whose total exposure time is the shortest possible? (3) How can we compute such sequences automatically for a specific camera, depth of field, and exposure level? (4) Finally, how do we convert the captured sequence into a single photo with the specified depth of field and exposure level?

Little is known about how to gather light efficiently from a specified DOF. Research on computational photography has not investigated the light-gathering ability of existing methods, and

has not considered the problem of optimizing exposure time for a desired DOF and exposure level. For example, even though there has been great interest in manipulating a camera's DOF through optical [28, 36, 69, 96, 115, 138] or computational [14, 29, 53, 54, 58, 75, 83] means, current approaches do so without regard to exposure time—they simply assume that the shutter remains open as long as necessary to reach the desired exposure level. This assumption is also used for high-dynamic range photography [31, 54], where the shutter must remain open for long periods in order to capture low-radiance regions in a scene. In contrast, here we capture photos with camera settings that are carefully chosen to minimize total exposure time for the desired DOF and exposure level.

Since shorter total exposure times reduce motion blur, our work can be thought of as complementary to recent *synthetic shutter* approaches whose goal is to reduce such blur. Instead of controlling aperture and focus, these techniques divide a given exposure interval into several shorter ones, with the same total exposure (*e.g.*, $n$ photos, each with $1/n$ the exposure time [113]; two photos, one with long and one with short exposure [131]; or one photo where the shutter opens and closes intermittently during the exposure [96]). These techniques do not increase light efficiency and do not rely on any camera controls other than the shutter. As such, they can be readily combined with our work, to confer the advantages of both methods.

The final step in light-efficient photography involves merging the captured photos to create a new one (Fig. 5.1). As such, our work is related to the well-known technique of extended-depth-of-field imaging. This technique creates a new photo whose DOF is the union of DOFs in a sequence, and has found wide use in microscopy [75], macro photography [10, 85] and photo manipulation [10, 85]. Current work on the subject concentrates on the problems of image merging [10, 87] and 3D reconstruction [75], and indeed we use an existing implementation [10] for our own merging step. However, the problem of how to best acquire such sequences remains open. In particular, the idea of controlling aperture and focus to optimize total exposure time has not been explored.

Our work offers four contributions over the state of the art. First, we develop a theory that leads to provably-efficient light-gathering strategies, and applies both to off-the-shelf cameras and to advanced camera designs [96, 113] under typical photography conditions. Second, from a practical standpoint, our analysis shows that the optimal (or near-optimal) strategies are very simple: for example, in the continuous case, a strategy that uses the widest-possible aperture for all photos is either globally optimal or it is very close to it (in a quantifiable sense). Third, our experiments with real scenes suggest that it is possible to compute good-quality synthesized photos using readily-available algorithms. Fourth, we show that despite requiring less total exposure

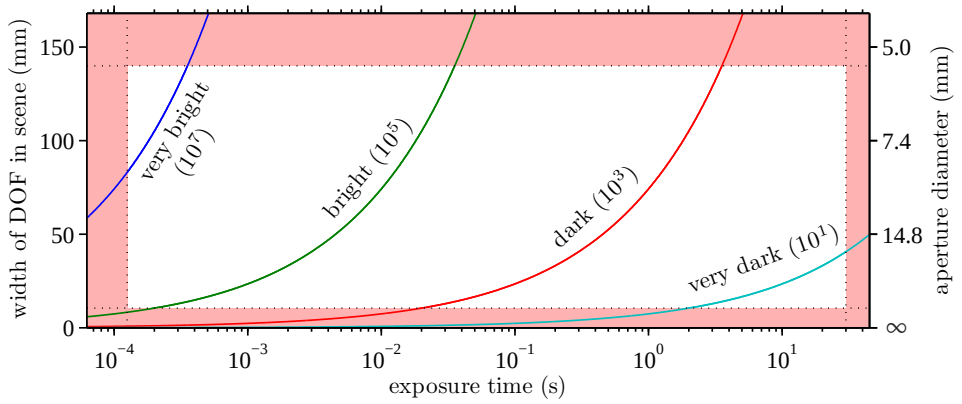



**Figure 5.2:** Each curve represents all pairs $(\tau, D)$ for which $\tau D^2 = L^*$ in a specific scene. Shaded zones correspond to pairs outside the camera limits (valid settings were $\tau \in [1/8000\,\text{s}, 30\,\text{s}]$ and $D \in [\text{f}/16, \text{f}/1.2]$ with $f = 85\,\text{mm}$). Also shown is the DOF corresponding to each diameter $D$. The maximum acceptable blur was set to $c = 25\,\mu\text{m}$, or about 3 pixels in our camera. Different curves represent scenes with different average radiance (relative units shown in brackets).

time than a single narrow-aperture shot, light-efficient photography provides more information about the scene (*i.e.* depth) and allows post-capture control of aperture and focus.

## 5.2 The Exposure Time *vs.* Depth of Field Tradeoff

The *exposure level* of a photo is the total radiant energy integrated by the camera's entire sensor while the shutter is open. The exposure level can influence significantly the quality of a captured photo because when there is no saturation or thermal noise, a pixel's signal-to-noise ratio (SNR) always increases with higher exposure levels.[1] For this reason, most modern cameras can automate the task of choosing an exposure level that provides high SNR for most pixels and causes little or no saturation.

Lens-based camera systems provide only two ways to control exposure level— the diameter of their aperture and the exposure time. We assume that all light passing through the aperture will reach the sensor plane, and that the average irradiance measured over this aperture is independent of the aperture's diameter. In this case, the exposure level $L$ satisfies

$$L \propto \tau D^2 , \tag{5.1}$$

where $\tau$ is exposure time and $D$ is the aperture diameter.

Now suppose that we have chosen a desired exposure level $L^*$. How can we capture a photo at

[1] Thermal effects, such as dark-current noise, become significant only for exposure times longer than a few seconds [56].

this exposure level? Eq. (5.1) suggests that there are only two general strategies for doing this—either choose a long exposure time and a small aperture diameter, or choose a large aperture diameter and a short exposure time. Unfortunately, both strategies have important side-effects: increasing exposure time can introduce motion blur when we photograph moving scenes [113, 131]; opening the lens aperture, on the other hand, affects the photo's *depth of field (DOF)*, *i.e.*, the range of distances where scene points do not appear out of focus. These side-effects lead to an important tradeoff between a photo's exposure time and its depth of field (Fig. 5.2):

> **Exposure Time *vs.* Depth of field Tradeoff:** *We can either achieve a desired exposure level $L^*$ with short exposure times and a narrow DOF, or with long exposure times and a wide DOF.*

In practice, the exposure time *vs.* DOF tradeoff limits the range of scenes that can be photographed at a given exposure level (Fig. 5.2). This range depends on scene radiance, the physical limits of the camera (*i.e.*, range of possible apertures and shutter speeds), as well as subjective factors (*i.e.*, acceptable levels of motion blur and defocus blur).

Our goal is to "break" this tradeoff by seeking novel photo acquisition strategies that capture a given depth of field at the desired exposure level $L^*$ much faster than traditional optics would predict. We briefly describe below the basic geometry and relations governing a photo's depth of field, as they are particularly important for our analysis.

### 5.2.1 Depth of Field Geometry

We assume that focus and defocus obey the standard thin lens model [92, 105]. This model relates three positive quantities (Eq. (5.2) in Table 5.1): the focus setting $v$, defined as the distance from the sensor plane to the lens; the distance $d$ from the lens to the in-focus scene plane; and the focal length $f$, representing the "focusing power" of the lens.

Apart from the idealized pinhole, all apertures induce spatially-varying amounts of defocus for points in the scene (Fig. 5.3a). If the lens focus setting is $v$, all points at distance $d$ from the lens will be in-focus. A scene point at distance $d' \neq d$, however, will be defocused: its image will be a circle on the sensor plane whose diameter $\sigma$ is called the *blur diameter*. For any given distance $d$, the thin-lens model tells us exactly what focus setting we should use to bring the plane at distance $d$ into focus, and what the blur diameter will be for points away from this plane (Eqs. (5.3) and (5.4), respectively).

For a given aperture and focus setting, the *depth of field* is the interval of distances in the scene whose blur diameter is below a maximum acceptable size $c$ (Fig. 5.3b).

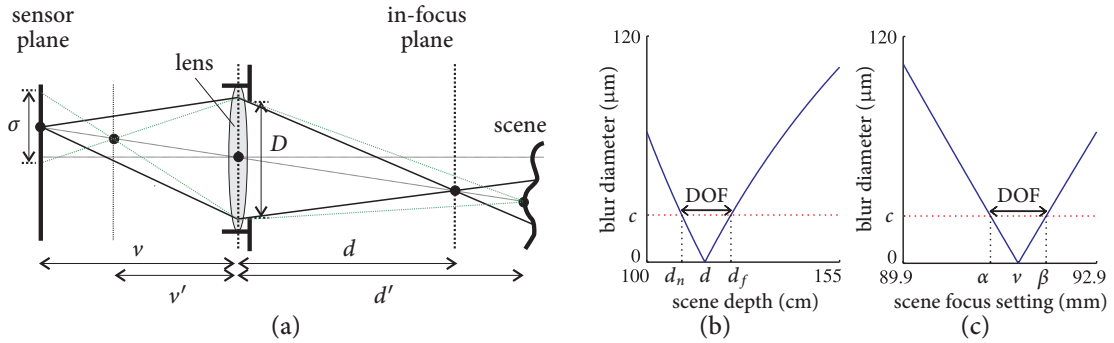



**Figure 5.3:** (a) Blur geometry for a thin lens. (b) Blur diameter as a function of distance to a scene point. The plot is for a lens with $F = 85\,\text{mm}$, focused at 117 cm with an aperture diameter of 5.31 mm (*i.e.*, an f/16 aperture in photography terminology). (c) Blur diameter and DOF represented in the space of focus settings.

| | | |
|---|---|---|
| thin lens law | $\frac{1}{v} + \frac{1}{d} = \frac{1}{F}$ | (5.2) |
| focus setting for distance $d$ | $v = \frac{dF}{d - F}$ | (5.3) |
| blur diameter for out-of-focus distance $d'$ | $\sigma = D\,\frac{F\lvert d' - d\rvert}{(d - F)\,d'}$ | (5.4) |
| aperture diameter whose DOF is interval $[\alpha, \beta]$ | $D = c\,\frac{\beta + \alpha}{\beta - \alpha}$ | (5.5) |
| focus setting whose DOF is interval $[\alpha, \beta]$ | $v = \frac{2\,\alpha\,\beta}{\alpha + \beta}$ | (5.6) |
| DOF endpoints for aperture diameter $D$ and focus $v$ | $\alpha, \beta = \frac{Dv}{D \pm c}$ | (5.7) |

**Table 5.1:** Basic equations governing focus and DOFs for the thin-lens model.

Since every distance in the scene corresponds to a unique focus setting (Eq. (5.3)), every DOF can also be expressed as an interval $[\alpha, \beta]$ in the space of focus settings. This alternate DOF representation gives us especially simple relations for the aperture and focus setting that produce a given DOF (Eqs. (5.5) and (5.6)) and, conversely, for the DOF produced by a given aperture and focus setting (Eq. (5.7)). We adopt this DOF representation for the rest of our analysis (Fig. 5.3c).

A key property of the depth of field is that it shrinks when the aperture diameter increases: from Eq. (5.4) it follows that for a given out-of-focus distance, larger apertures always produce

larger blur diameters. This equation is the root cause of the exposure time *vs.* depth of field tradeoff.

## 5.3 The Synthetic DOF Advantage

Suppose that we want to capture a single photo with a specific exposure level $L^*$ *and* a specific depth of field $[\alpha, \beta]$. How quickly can we capture this photo? The basic DOF geometry of Sec. 5.2.1 tells us we have no choice: there is only one aperture diameter that can span the given depth of field (Eq. (5.5)), and only one exposure time that can achieve a given exposure level with that diameter (Eq. (5.1)). This exposure time is[2]

$$\tau^{one} = L^* \cdot \left( \frac{\beta - \alpha}{c\,(\beta + \alpha)} \right)^2 . \tag{5.8}$$

The key idea of our approach is that while lens optics do not allow us to reduce this time without compromising the DOF or the exposure level, we *can* reduce it by taking more photos. This is based on a simple observation that takes advantage of the different rates at which exposure time and DOF change: if we increase the aperture diameter and adjust exposure time to maintain a constant exposure level, its DOF shrinks (at a rate of about $1/D$), but the exposure time shrinks much faster (at a rate of $1/D^2$). This opens the possibility of "breaking" the exposure time *vs.* DOF tradeoff by capturing a sequence of photos that jointly span the DOF in less total time than $\tau^{one}$ (Fig. 5.1).

Our goal is to study this idea in its full generality, by finding capture strategies that are provably time-optimal. We therefore start from first principles, by formally defining the notion of a *capture sequence* and of its *synthetic depth of field*:

**Definition 1** (Photo Tuple). *A tuple* $\langle D, \tau, \nu \rangle$ *that specifies a photo's aperture diameter, exposure time, and focus setting, respectively.*

**Definition 2** (Capture Sequence). *A finite ordered sequence of photo tuples.*

**Definition 3** (Synthetic Depth of Field). *The union of DOFs of all photo tuples in a capture sequence.*

We will use two efficiency measures: the *total exposure time* of a sequence is the sum of the exposure times of all its photos; the *total capture time*, on the other hand, is the actual time

[2] The apertures and exposure times of real cameras span finite intervals and, in many cases, take discrete values. Hence, in practice, Eq. (5.8) holds only approximately.

it takes to capture the photos with a specific camera. This time is equal to the total exposure time, plus any overhead caused by camera internals (computational and mechanical). We now consider the following general problem:

> **Light-Efficient Photography**: *Given a set $\mathcal{D}$ of available aperture diameters, construct a capture sequence such that: (1) its synthetic DOF is equal to $[\alpha, \beta]$; (2) all its photos have exposure level $L^*$; (3) the total exposure time (or capture time) is smaller than $\tau^{one}$; and (4) this time is a global minimum over all finite capture sequences.*

Intuitively, whenever such a capture sequence exists, it can be thought of as being optimally more efficient than single-shot photography in gathering light. Below we analyze three instances of the light-efficient photography problem. In all cases, we assume that the exposure level $L^*$, depth of field $[\alpha, \beta]$, and aperture set $\mathcal{D}$ are known and fixed.

**Noise and Quantization Properties.** Because we hold exposure level constant our analysis already accounts for noise implicitly. This follows from the fact that most sources of noise (photon noise, sensor noise, and quantization noise) depend only on exposure level. The only exception is thermal or dark-current noise, which increases with exposure time [56]. Therefore, all photos we consider have similar noise properties, except for thermal noise, which will be lower for light-efficient sequences because they involve shorter exposure times.

Another consequence of holding exposure level constant is that all photos we consider have the same dynamic range, since all photos are exposed to the same brightness, and have similar noise properties for quantization. Therefore, standard techniques for HDR imaging [31, 78] are complementary to our analysis, since we can apply light-efficient capture for each exposure level in an HDR sequence.

## 5.4 Theory of Light-Efficient Photography

### 5.4.1 Continuously-Variable Aperture Diameters

Many manual-focus SLR lenses as well as programmable-aperture systems [138] allow their aperture diameter to vary continuously within some interval $\mathcal{D} = [D_{min}, D_{max}]$. In this case, we prove that the optimal capture sequence has an especially simple form—it is unique, it uses the same aperture diameter for all tuples, and this diameter is either the maximum possible or a diameter close to that maximum.

More specifically, consider the following special class of capture sequences:

**Definition 4** (Sequences with Sequential DOFs). *A capture sequence has sequential DOFs if for every pair of adjacent photo tuples, the right endpoint of the first tuple's DOF is the left endpoint of the second.*

The following theorem states that the solution to the light-efficient photography problem is a specific sequence from this class:

**Theorem 1** (Optimal Capture Sequence for Continuous Apertures). *(1) If the DOF endpoints satisfy $\beta < (7 + 4\sqrt{3})\alpha$, the sequence that globally minimizes total exposure time is a sequence with sequential DOFs whose tuples all have the same aperture. (2) Define $D(k)$ and $n$ as follows:*

$$D(k) = c\,\frac{\sqrt[k]{\beta} + \sqrt[k]{\alpha}}{\sqrt[k]{\beta} - \sqrt[k]{\alpha}}\,, \qquad n = \left\lfloor \frac{\log \frac{\alpha}{\beta}}{\log\left(\frac{D_{max}-c}{D_{max}+c}\right)} \right\rfloor\,. \tag{5.9}$$

*The aperture diameter $D^*$ and length $n^*$ of the optimal sequence is given by*

$$D^* = \begin{cases} D(n) & \text{if } \frac{D(n)}{D_{max}} > \sqrt{\frac{n}{n+1}} \\ D_{max} & \text{otherwise.} \end{cases} \qquad n^* = \begin{cases} n & \text{if } \frac{D(n)}{D_{max}} > \sqrt{\frac{n}{n+1}} \\ n+1 & \text{otherwise.} \end{cases}\,. \tag{5.10}$$

Theorem 1 specifies the optimal sequence indirectly, via a "recipe" for calculating the optimal length and the optimal aperture diameter (Eqs. (5.9) and (5.10)). Informally, this calculation involves three steps. The first step defines the quantity $D(k)$; in our proof of Theorem 1 (see Appendix D), we show that this quantity represents the only aperture diameter that can be used to "tile" the interval $[\alpha, \beta]$ with exactly $k$ photo tuples of the same aperture. The second step defines the quantity $n$; in our proof, we show that this represents the largest number of photos we can use to tile the interval $[\alpha, \beta]$ with photo tuples of the same aperture. The third step involves choosing between two "candidates" for the optimal solution—one with $n$ tuples and one with $n + 1$.

Theorem 1 makes explicit the somewhat counter-intuitive fact that the most light-efficient way to span a given DOF $[\alpha, \beta]$ is to use images whose DOFs are very narrow. This fact applies broadly, because Theorem 1's inequality condition for $\alpha$ and $\beta$ is satisfied for all lenses for consumer photography that we are aware of (*e.g.*, see [2]).[3] See Figs. 5.4 and 5.5 for an application of this theorem to a practical example.

[3] To violate the condition, a lens must have an extremely short minimum focusing distance of under $1.077f$. The condition can still hold for macro lenses with a stated minimum focusing distance of 0, since this is measured relative to the front-most glass surface, and the effective lens center is deeper inside.

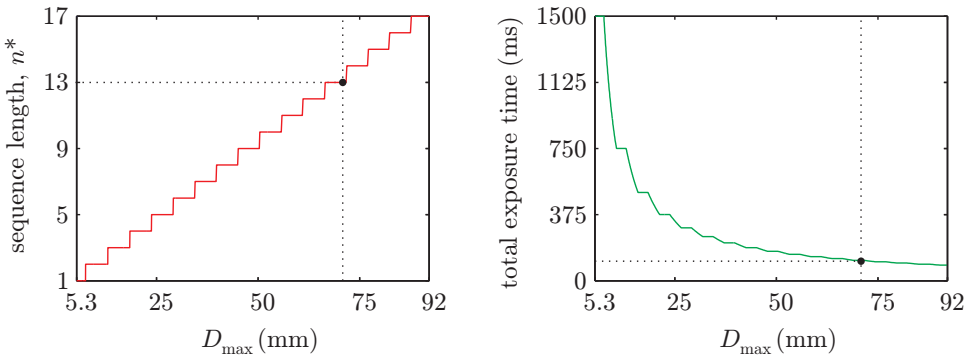



**Figure 5.4:** Optimal light-efficient photography of a "dark" subject using a lens with a continuously-variable aperture ($F$ = 85 mm). To cover the DOF ([110 cm, 124 cm]) in a single photo, we need a long 1.5$s$ exposure to achieve the desired exposure level. Together, the two graphs specify the optimal capture sequences when the aperture diameter is restricted to the range $[\text{f/16}, D_{\max}]$; for each value of $D_{\max}$, Theorem 1 gives a unique optimal sequence. As $D_{\max}$ increases, the number of photos (left) in the optimal sequence increases, and the total exposure time (right) of the optimal sequence falls dramatically. The dashed lines show that when the maximum aperture is f/1.2 (71 mm), the optimal synthetic DOF consists of $n^*$ = 13 photos (corresponding to $D^*$ = 69.1 mm), which provides a speedup of 13× over single-shot photography.

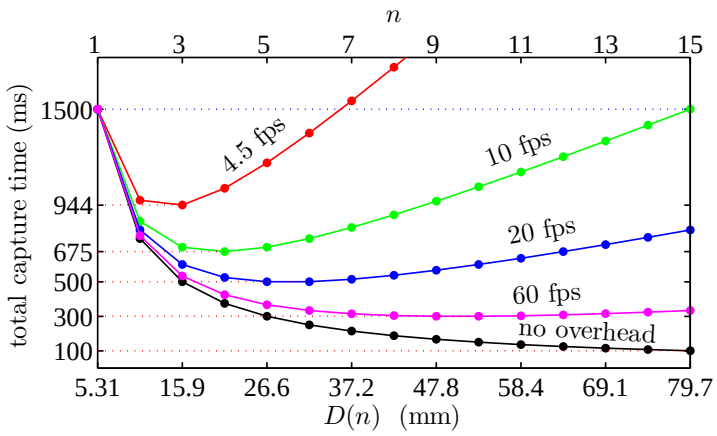



**Figure 5.5:** The effect of camera overhead for various frame-per-second (fps) rates. Each point in the graphs represents the total capture time of a sequence that spans the DOF and whose photos all use the diameter $D(n)$ indicated. Even though overhead reduces the efficiency of long sequences, capturing synthetic DOFs is faster than single-shot photography even for low-fps rates; for current off-the-shelf cameras with high-fps rates, the speedups can be very significant.

Note that Theorem 1 specifies the number of tuples in the optimal sequence and their aperture diameter, but does not specify their exposure times or focus settings. The following lemma shows that specifying those quantities is not necessary because they are determined uniquely. Importantly, Lemma 1 gives us a recursive formula for computing the exposure time and focus setting of each tuple in the sequence:

**Lemma 1** (Construction of Sequences with Sequential DOFs). *Given a left DOF endpoint* $\alpha$,

*every ordered sequence $D_1, \ldots, D_n$ of aperture diameters defines a unique capture sequence with sequential DOFs whose n tuples are*

$$\left\langle\, D_i\,,\ \frac{L^*}{{D_i}^2}\,,\ \frac{D_i + c}{D_i}\,\alpha_i \,\right\rangle\ ,\quad i\ =\ 1, \ldots, n\ , \tag{5.11}$$

*with $\alpha_i$ given by the following recursive relation:*

$$\alpha_i\ =\ \begin{cases} \alpha & \text{if } i = 1\ , \\ \frac{D_i + c}{D_i - c}\,\alpha_{i-1} & \text{otherwise.} \end{cases} \tag{5.12}$$

### 5.4.2 Discrete Aperture Diameters

Modern auto-focus lenses often restrict the aperture diameter to a discrete set of choices, $\mathcal{D} = \{D_1, \ldots, D_m\}$. These diameters form a geometric progression, spaced so that the aperture area doubles every two or three steps. Unlike the continuous case, the optimal capture sequence is not unique and may contain several distinct aperture diameters. To find an optimal sequence, we reduce the problem to integer linear programming [86]:

**Theorem 2** (Optimal Capture Sequence for Discrete Apertures)**.** *There exists an optimal capture sequence with sequential DOFs whose tuples have a non-decreasing sequence of aperture diameters. Moreover, if $n_i$ is the number of times diameter $D_i$ appears in the sequence, the multiplicities $n_1, \ldots, n_m$ satisfy the integer program*

$$\text{minimize}\quad \textstyle\sum_{i=1}^{m} n_i\,\frac{L^*}{{D_i}^2} \tag{5.13}$$

$$\text{subject to}\quad \textstyle\sum_{i=1}^{m} n_i \log\frac{D_i - c}{D_i + c}\ \leq\ \log\frac{\alpha}{\beta} \tag{5.14}$$

$$n_i \geq 0 \tag{5.15}$$

$$n_i \text{ integer}\ . \tag{5.16}$$

See Appendix D for a proof. As with Theorem 1, Theorem 2 does not specify the focus settings in the optimal capture sequence. We use Lemma 1 for this purpose, which explicitly constructs it from the apertures and their multiplicities.

While it is not possible to obtain a closed-form expression for the optimal sequence, solving the integer program for any desired DOF is straightforward. We use a simple branch-and-bound method based on successive relaxations to linear programming [86]. Moreover, since the optimal sequence depends only on the relative DOF size $\frac{\alpha}{\beta}$, we pre-compute it exactly for all relative

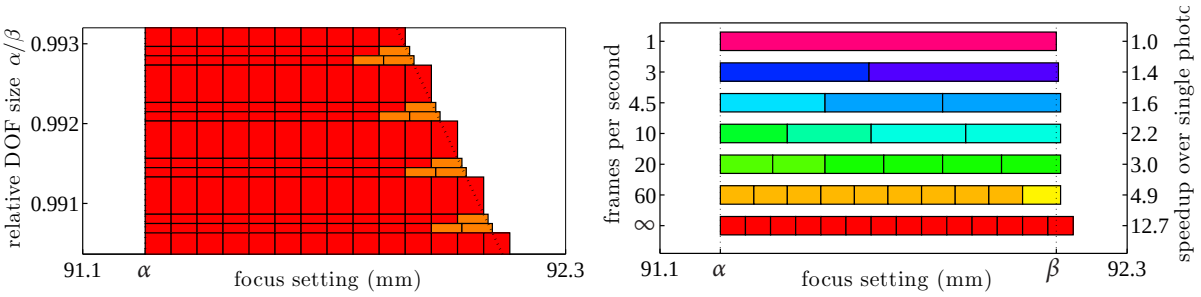



**Figure 5.6:** Optimal light-efficient photography with discrete apertures, shown for a Canon EF85mm 1.2L lens (23 apertures, illustrated in different colors). (a) For a depth of field whose left endpoint is $\alpha$, we show optimal capture sequences for a range of relative DOF sizes $\frac{\alpha}{\beta}$. These sequences can be read horizontally, with subintervals corresponding to the apertures determined by Theorem 2. Note that when the DOF is large, they effectively approximate the continuous case. The diagonal dotted line shows the minimum DOF to be spanned. (b) Visualizing the optimal capture sequence as a function of the camera overhead for the DOF $[\alpha, \beta]$. Note that as the overhead increases, the optimal sequence involves fewer photos with larger DOFs (*i.e.*, smaller apertures).

sizes and store it in a lookup table (Fig. 5.6a).

### 5.4.3 Discrete Aperture Diameters Plus Overhead

Our treatment of discrete apertures generalizes easily to account for camera overhead. We model overhead as a per-shot constant, $\tau^{over}$, that expresses the minimum delay between the time that the shutter closes and the time it is ready to open again for the next photo. To find the optimal sequence, we modify the objective function of Theorem 2 so that it measures for total capture time rather than total exposure time:

$$\textit{minimize} \quad \sum_{i=1}^{m} n_i \left[ \tau^{over} + \frac{L^*}{{D_i}^2} \right] . \tag{5.17}$$

Clearly, a non-negligible overhead penalizes long capture sequences and reduces the synthetic DOF advantage. Despite this, Fig. 5.6b shows that synthetic DOFs offer significant speedups even for current off-the-shelf cameras. These speedups will be amplified further as camera manufacturers continue to improve their frames-per-second rate.

## 5.5 Depth of Field Compositing and Resynthesis

While each light-efficient sequence captures a synthetic DOF, merging the input photos into a single photo with the desired DOF requires further processing. To achieve this, we use an existing depth-from-focus and compositing technique [10], and propose a simple extension that

allows us to reshape the DOF, to synthesize photos with new camera settings as well.

**DOF Compositing.** To reproduce the desired DOF, we adopted the Photomontage method [10] with default parameters, which is based on maximizing a simple "focus measure" that evaluates local contrast according to the difference-of-Gaussians filter. In this method, each pixel in the composite has a label that indicates the input photo for which the pixel is in-focus. The pixel labels are then optimized using a Markov random field network that is biased toward piecewise smoothness [25]. Importantly, the resulting composite is computed as a blend of photos in the gradient domain, which reduces artifacts at label boundaries, including those due to misregistration.

**3D Reconstruction.** The DOF compositing operation produces a coarse depth map as an intermediate step. This is because labels correspond to input photos, and each input photo defines an in-focus depth according to the focus setting with which it was captured. As our results show, this coarse depth map is sufficient for good-quality resynthesis (Figs. 5.9–5.7, 5.8). For greater depth accuracy, particularly when the capture sequence consists of only a few photos, we can apply more sophisticated depth-from-defocus analysis, *e.g.*, [120], that reconstructs depth by modeling how defocus varies over the whole sequence.

**Synthesizing Photos for Novel Focus Settings and Aperture Diameters.** To synthesize novel photos with different camera settings, we generalize DOF compositing and take advantage of the different levels of defocus throughout the capture sequence. Intuitively, rather than selecting pixels at in-focus depths from the input sequence, we use the recovered depth map to select pixels with appropriate levels of defocus according to the desired synthetic camera setting.

We proceed in four basic steps. First, given a specific focus and aperture setting, we use Eq. (5.4) and the coarse depth map to assign a blur diameter to each pixel in the final composite. Second, we use Eq. (5.4) again to determine, for each pixel in the composite, the input photo whose blur diameter that corresponds to the pixel's depth matches most closely. Third, for each depth layer, we synthesize a photo with the novel focus and aperture setting, under the assumption that the entire scene is at that depth. To do this, we use the blur diameter for this depth to define an interpolation between two of the input photos. Fourth, we generate the final composite by merging all these synthesized images into one photo using the same gradient-domain blending as in DOF compositing, and using the same depth labels.[4]

[4]Note that given a blur diameter there are two possible depths that correspond to it, one on each side of the

To interpolate between the input photos we currently use simple linear cross-fading, which we found to be adequate when the DOF is sampled densely enough (*i.e.*, with 5 or more images). For greater accuracy when fewer input images are available, more computationally intensive frequency-based interpolation [29] could also be used. Note that blur diameter can also be extrapolated, by synthetically applying the required additional blur. As discussed in Sec. 4.8, there are limitations to this extrapolation. While extrapolated wider apertures can model the resulting increase in defocus, we have limited ability to reduce the DOF for an input image, which would entail decomposing an in-focus region into finer depth gradations [99].

## 5.6 Results and Discussion

To evaluate our technique we show results and timings for experiments performed with two different cameras—a high-end digital SLR and a compact digital camera. All photos were captured at the same exposure level for each experiment, determined by the camera's built-in light meter. In each case, we captured (1) a narrow-aperture photo, which serves as ground truth, and (2) the optimal capture sequence for the equivalent DOF.[5]

The digital SLR we used was the Canon EOS-1Ds Mark II (HAMSTER and FACE datasets) with a wide-angle fixed focal length lens (Canon EF85mm 1.2L). We operated the camera at its highest resolution of 16 MP (4992 × 3328) in RAW mode. To define the desired DOF, we captured a narrow-aperture photo using an aperture of f/16. For both datasets, the DOF we used was [98 cm, 108 cm], near the minimum focusing distance of the lens, and the narrow-aperture photo required an exposure time of 800 ms.

The compact digital camera we used was the Canon S3 IS, at its widest-angle zoom setting with a focal length of 6 mm (SIMPSONS dataset). We used the camera to record 2 MP (1600 × 1200 pixels) JPEG images. To define the desired DOF, we captured a photo with the narrowest aperture of f/8. The DOF we used was [30 cm, 70 cm], and the narrow-aperture photo required an exposure time of 500 ms.

- **HAMSTER dataset** Still life of a hamster figurine (16 cm tall), posed on a table with various other small objects (Fig. 5.7). The DOF covers the hamster and all the small objects, but not the background composed of cardboard packing material.
- **FACE dataset** Studio-style 2/3 facial portrait of a subject wearing glasses, resting his chin on his hands (Fig. 5.8). The DOF extends over the subject's face and the left side of the

focus plane (Fig. 5.3b, Sec. 2.7). We resolve this by choosing the matching input photo whose focus setting is closest to the synthetic focus setting.

[5] For additional results and videos, see `http://www.cs.toronto.edu/~hasinoff/lightefficient/`.

body closest the camera.

- **Simpsons dataset** Near-macro sequence of a messy desk (close objects magnified 1:5), covered in books, papers, and tea paraphernalia, on top of which several plastic figurines have been arranged (Fig. 5.9). The DOF extends from red tea canister to the pale green book in the background.

**Implementation details.** To compensate for the distortions that occur with changes in focus setting, we align the photos according to a one-time calibration method that fits a simplified radial magnification model to focus setting [127].

We determined the maximum acceptable blur diameter, $c$, for each camera by qualitatively assessing focus using a resolution chart. The values we used, 25 μm (3.5 pixels) and 5 μm (1.4 pixels) for the digital SLR and compact camera respectively, agree with the standard values cited for sensors of those sizes [105].

To process the 16 MP synthetic DOFs captured with the digital SLR more efficiently, we divided the input photos into tiles of approximately 2 MP each, so that all computation could take place in main memory. To improve continuity at tile boundaries, we use tiles that overlap with their neighbors by 100 pixels. Even so, as Fig. 5.8d illustrates, merging per-tile results that were computed independently can introduce depth artifacts along tile boundaries. In practice, these tile-based artifacts do not pose problems for resynthesis, because they are restricted to textureless regions, for which realistic resynthesis does not depend on accurate depth assignment.

**Timing comparisons and optimal capture sequences.** To determine the optimal capture sequences, we assumed zero camera overhead and applied Theorem 2 for the chosen DOF and exposure level, according to the specifications of each camera and lens. The optimal sequences involved spanning the DOF using the largest aperture in both cases. As Figs. 5.7–5.9 show, these sequences led to significant speedups in exposure time—11.9× and 2.5× for our digital SLR and compact digital camera respectively.[6]

For a hypothetical camera overhead of 17 ms (corresponding to a 60 fps camera), the optimal capture sequence satisfies Eq. (5.17), which changes the optimal strategy for the digital SLR only (Hamster and Face datasets). At this level of overhead, the optimal sequence for this case takes 220 ms to capture[7], compared to 800 ms for one narrow-aperture photo. This reduces the

[6]By comparison, the effective speedup provided by optical image stabilization for hand-held photography is 8–16×, when the scene is static. Gains from light efficient photography are complementary to such improvements in lens design.

[7]More specifically, the optimal sequence involves spanning the DOF with 7 photos instead of 14. This sequence consists of 1 photo captured at f/2, plus 3 photos each at f/2.2 and f/2.5.

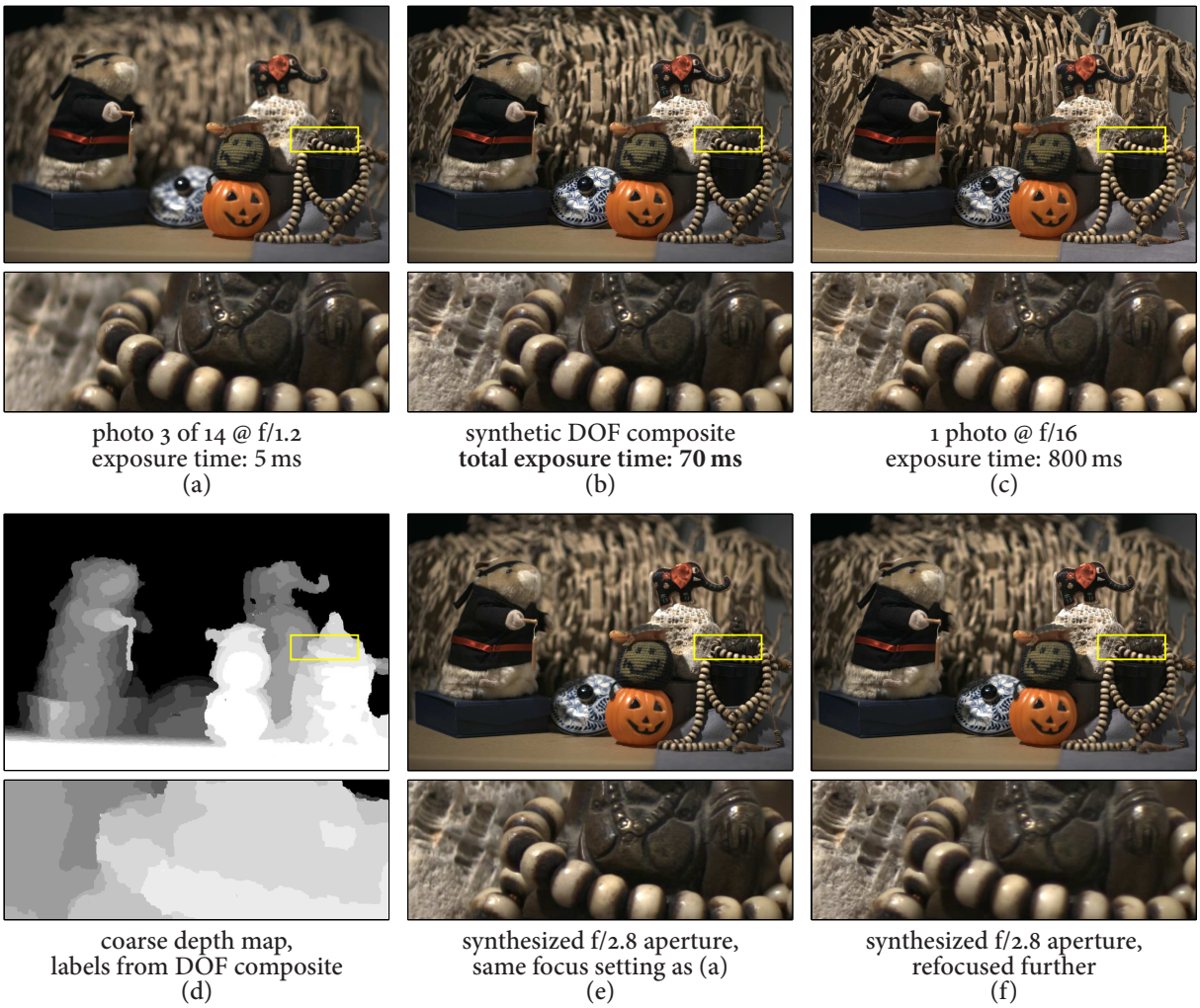



**Figure 5.7:** HAMSTER dataset. Light efficient photography timings and synthesis, for several real scenes, captured using a compact digital camera and a digital SLR. (a) Sample wide-aperture photo from the synthetic DOF sequence. (b) DOF composites synthesized from this sequence. (c) Narrow-aperture photos spanning an equivalent DOF, but with much longer exposure time. (d) Coarse depth map, computed from the labeling we used to compute (b). (e) Synthetically changing aperture size, focused at the same setting as (a). (f) Synthetically changing focus setting as well, for the same synthetic aperture as (e).

speedup to 3.6×.

**DOF compositing.** Despite the fact that it relies on a coarse depth map, our compositing scheme is able to reproduce high-frequency detail over the whole DOF, without noticeable artifacts, even in the vicinity of depth discontinuities (Figs. 5.7b, 5.8b, and 5.9b). The narrow-aperture photos represent ground truth, and visually they are almost indistinguishable from our composites.

The worst compositing artifact occurs in the HAMSTER dataset, at the handle of the pumpkin

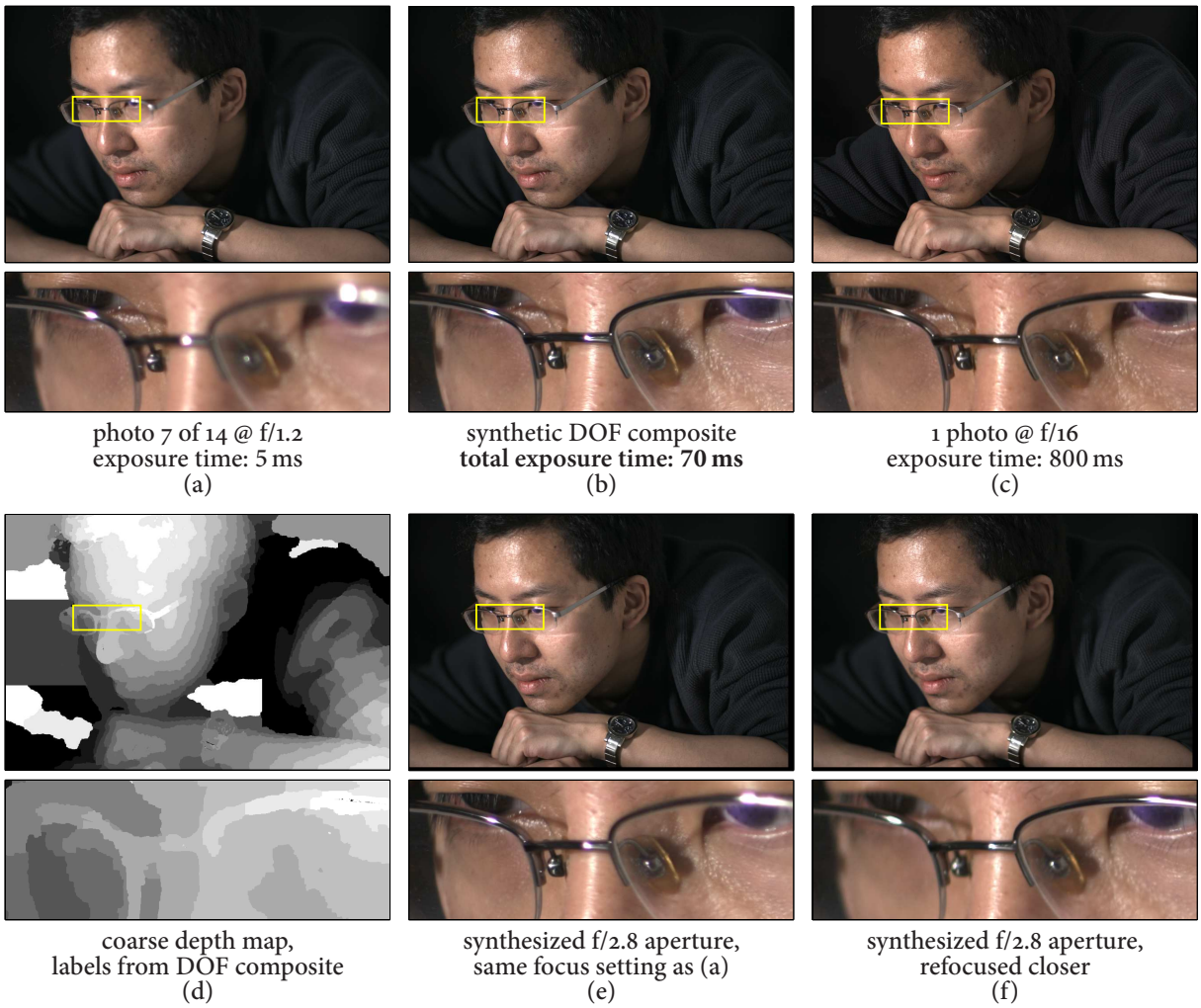



**Figure 5.8:** Face dataset. Light efficient photography timings and synthesis, for several real scenes, captured using a compact digital camera and a digital SLR. (a) Sample wide-aperture photo from the synthetic DOF sequence. (b) DOF composites synthesized from this sequence. (c) Narrow-aperture photos spanning an equivalent DOF, but with much longer exposure time. (d) Coarse depth map, computed from the labeling we used to compute (b). Tile-based processing leads to depth artifacts in low-texture regions, but these do not affect the quality of resynthesis. (e) Synthetically changing aperture size, focused at the same setting as (a). (f) Synthetically changing focus setting as well, for the same synthetic aperture as (e).

container, which is incorrectly assigned to a background depth (Fig. 5.10). This is an especially challenging region because the handle is thin and low-texture compared to the porcelain lid behind it.

Note that while the synthesized photos satisfy our goal of spanning a specific DOF, objects outside that DOF will appear more defocused than in the corresponding narrow-aperture photo. For example, the cardboard background in the Hamster dataset is not included in the DOF (Fig. 5.11). This background therefore appears slightly defocused in the narrow-aperture f/16

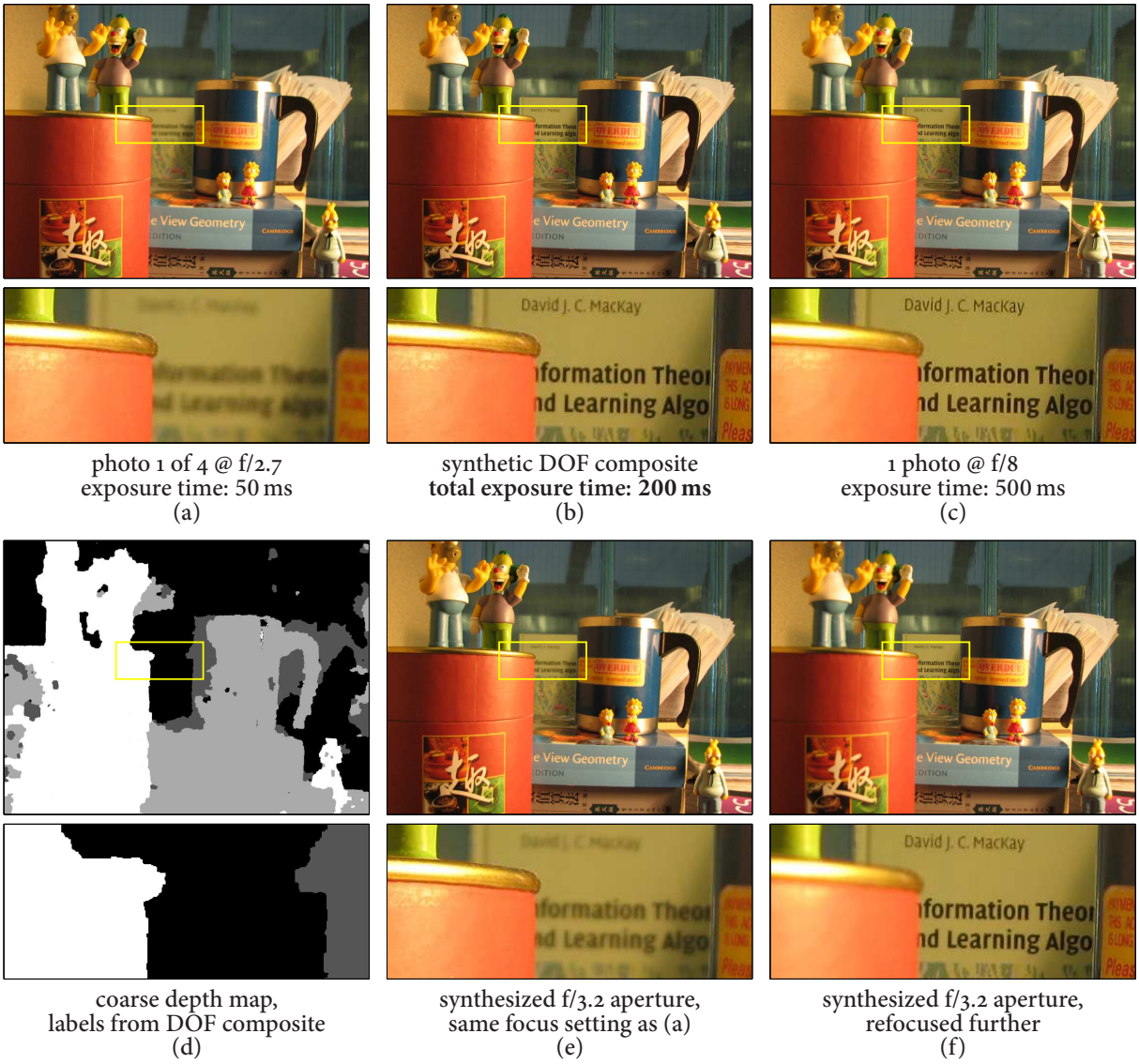



**Figure 5.9:** SIMPSONS dataset. Light efficient photography timings and synthesis, for several real scenes, captured using a compact digital camera and a digital SLR. (a) Sample wide-aperture photo from the synthetic DOF sequence. (b) DOF composites synthesized from this sequence. (c) Narrow-aperture photos spanning an equivalent DOF, but with much longer exposure time. (d) Coarse depth map, computed from the labeling we used to compute (b). (e) Synthetically changing aperture size, focused at the same setting as (a). (f) Synthetically changing focus setting as well, for the same synthetic aperture as (e).

photo, and strongly defocused in the synthetic DOF composite. This effect is expected, since outside the synthetic DOF, the blur diameter will increase proportional to the wider aperture diameter (Eq. (5.4)). For some applications, such as portrait photography, increased background defocus may be a beneficial feature.


**Figure 5.10:** Compositing failure for the Hamster dataset (Fig. 5.7). Elsewhere this scene is synthesized realistically. The depth-from-focus method employed by the Photomontage method breaks down at the handle of the pumpkin container, incorrectly assigning it to a background layer. This part of the scene is challenging to reconstruct because strong scene texture is visible "through" the defocused handle [42], whereas the handle itself is thin and low-texture.

**Depth maps and DOF compositing.** Despite being more efficient to capture, sequences with synthetic DOFs provide 3D shape information at no extra acquisition cost (Figs. 5.7d, 5.8d, and 5.9d). Using the method described in Sec. 5.5, we also show results of using this depth map to compute novel images whose aperture and focus setting was changed synthetically (Figs. 5.7e–f, 5.8e–f, and 5.9e–f). As a general rule, the more light-efficient a capture sequence is, the denser it is, and therefore the wider the range it offers for synthetic refocusing.

**Focus control and overhead.** Neither of our cameras provide the ability to control focus programmatically, so we used several methods to circumvent this limitation. For our digital SLR, we used a computer-controlled stepping motor to drive the lens focusing ring mechanically [4]. For our compact digital camera, we exploited modified firmware that provides general scripting capabilities [6]. Unfortunately, both these methods incur high additional overhead, effectively limiting us to about 1 fps.

Note that mechanical refocusing contributes relatively little overhead for the SLR, since ultrasonic lenses, like the Canon 85mm 1.2L we used, are fast. Our lens takes 3.5 ms to refocus from one photo in the sequence to the next, for a total of 45 ms to cover the largest possible DOF spanned by a single photo. In addition, refocusing can potentially be executed in parallel with other tasks such as processing the previous image. Such parallel execution already occurs in the Canon's "autofocus servo" mode, in which the camera refocuses continuously on a moving subject.

While light-efficient photography may not be practical using our current prototypes, it will


**Figure 5.11:** Background defocus for the HAMSTER dataset. Because the cardboard background lies outside the DOF, it is slightly defocused in the narrow-aperture photo. In the synthetic DOF composite, however, this background is defocused much more significantly. This effect is expected, because the synthetic DOF composite is created from much wider-aperture photos, and the blur diameter scales linearly with aperture. The synthetic DOF composite only produces in-focus images of objects lying *within* the DOF.

become increasingly so, as newer cameras begin to expose their focusing API directly and new CMOS sensors increases throughput. For example, the Canon EOS-1Ds Mark III provides remote focus control for all Canon EF lenses, and the recently released Casio EX-F1 can capture 60 fps at 6 MP. Even though light-efficient photography will benefit from the latest camera technology, as Fig. 5.5 shows, we can still realize time savings at slower frames-per-second rates.

**Handling motion in the capture sequence.** Because of the high overhead due to our focus control mechanisms, we observed scene motion in two of our capture sequences. The SIMPSONS dataset shows a subtle change in brightness above the green book in the background, because the person taking the photos moved during acquisition, casting a moving shadow on the wall. This is not an artifact and did not affect our processing. For the FACE dataset, the subject moved slightly during acquisition of the optimal capture sequence. To account for this motion, we performed a global rigid 2D alignment between successive images using Lucas-Kanade registration [19].

Despite this inter-frame motion, our approach for creating photos with a synthetic DOF (Sec. 5.5) generates results that are free of artifacts. In fact, the effects of this motion are only possible to see only in the videos that we create for varying synthetic aperture and focus settings. Specifically, while each still in the videos appears free of artifacts, successive stills contain a slight but noticeable amount of motion.

We emphasize the following two points. First, had we been able to exploit the internal focus control mechanism of the camera (a feature that newer cameras like the Canon EOS-1Ds

Mark III provide), the inter-frame motion for Face dataset would have been negligible, making the above registration step unnecessary. Second, even with fast internal focus control, residual motions would occur when photographing fast-moving subjects; our results in this sequence suggest that even in that case, our simple merging method should be sufficient to handle such motions with little or no image degradation.

## 5.7 Comparison to Alternative Camera Designs

While all the previous analysis for light-efficient capture assumed a conventional camera, it is instructive to compare our method to other approaches based on specially designed hardware. These approaches claim the ability to extend the DOF, which is analogous to reduced capture time in our formulation, since time savings can be applied to capture additional photos and extend the DOF.

**Light field cameras.** The basic idea of a light field camera is to trade sensor resolution for an increased number of viewpoints in a single photo [47, 85, 115]. Our approach is both more light-efficient and orthogonal to light field cameras, because despite being portrayed as such [85, 115], light field cameras *do not* have the ability to extend the DOF compared to regular wide-aperture photography. The authors of [85] have confirmed the following analysis [72].

First, consider a conventional camera with an $NK \times NK$ pixel sensor, whose aperture is set to the widest diameter of $D_{\max}$. For comparison, consider a light field camera built by placing an $N \times N$ lenslet array in front of the same sensor, yielding $N \times N$ reduced-resolution sub-images from $K^2$ different viewpoints [85]. Since each sub-image corresponds to a smaller effective aperture with diameter $D_{\max}/K$, the blur diameter for every scene point will be reduced by a factor of $K$ as well (Eq. (5.4)).

While the smaller blur diameters associated with the light field camera apparently serve to extend the DOF, this gain is misleading, because the sub-images have reduced resolution. By measuring blur diameter in pixels, we can see that an identical DOF "extension" can be obtained from a *regular* wide-aperture photo, just by resizing it by a factor of $1/K$ to match the sub-image resolution.

Indeed, an advantage of the light field camera is that by combining the sub-images captured from different viewpoints we can "refocus" the light field [61] by synthesizing reduced-resolution photos that actually have *reduced* DOF compared to regular photography. It is this reduction in DOF that allows us to refocus anywhere within the overall DOF defined by each sub-image,

which is the same as the DOF of the conventional camera.

Since the light field camera and regular wide-aperture photography collect the same number of photons, their noise properties are similar. In particular, both methods can benefit equally from noise reduction due to averaging, which occurs both when synthetically refocusing the light field [61] followed by compositing [10], and when resizing the regular wide-aperture image.

In practice, light field cameras are actually less light-efficient than wide-aperture photography, because they require stopping down the lens to avoid overlap between lenslet images [85], or they block light as a result of the imperfect packing of optical elements [47]. The above analysis also holds for the heterodyne light field camera [115], where the mask placed near the sensor blocks 70 % of the light, except that the sub-images are defined in frequency space.

**Wavefront coding.** Wavefront coding methods rely on a special optical element that effectively spreads defocus evenly over a larger DOF, and then recovers the underlying in-focus image using deconvolution [28]. While this approach is powerful, it exploits a tradeoff that is also orthogonal to our analysis. Wavefront coding can extend perceived DOF by a factor of $K = 2$ to 10, but it suffers from reduced SNR, especially at high frequencies [28], and it provides no 3D information. The need to deconvolve the image is another possible source of error when using wavefront coding, particularly since the point-spread function is only approximately constant over the extended DOF.

To compare wavefront coding with our approach in a fair way, we fix the total exposure time, $\tau$ (thereby collecting the same number of photons), and examine the SNR of the restored in-focus photos. Roughly speaking, wavefront coding can be thought of as capturing a single photo while sweeping focus through the DOF [55]. By contrast, our approach involves capturing $K$ in-focus photos spanning the DOF, each allocated exposure time of $\tau/K$. The sweeping analogy suggests that wavefront coding can do no better than our method in terms of SNR, because it collects the same number of "in-focus" photons for a scene at a given depth.

**Aperture masks.** Narrow apertures on a conventional camera can be thought of as masks in front of the widest aperture, however it is possible to block the aperture using more general masks as well. For example, ring-shaped apertures [88, 123] have a long history in astronomy and microscopy, and recent methods have proposed using coded binary masks in conjunction with regular lenses [69, 115]. Note that the use of aperture masks is complementary to our analysis, in that however much a particular mask shape can effectively extend the DOF, our analysis suggests that this mask should be scaled to be used with large apertures.

While previous analysis suggests that ring-shaped apertures yield no light-efficient benefit [123], the case for coded aperture masks is less clear, despite recent preliminary analysis that suggests the same [70]. The advantage of coded masks is their ability to preserve high frequencies that would otherwise be lost to defocus, so the key question is whether coded apertures increase effective DOF enough to justify blocking about 50 % of the light.

Resolving the light-efficiency of aperture masks requires a more sophisticated error analysis of the in-focus reconstruction, going beyond the geometric approach to DOF. We develop such a framework in the following chapter, as Levin, *et al.* [70] have also done independently. Unlike the wavefront coding case, this analysis is complicated by the fact that processing a coded-aperture image depends on the non-trivial task of depth recovery, which determines the spatially-varying deconvolution needed to reconstruct the in-focus image.

## Summary

In this chapter we studied the use of dense, wide-aperture photo sequences as a light-efficient alternative to single-shot, narrow-aperture photography. While our emphasis has been on the underlying theory, we believe that our results will become increasingly relevant as newer, off-the-shelf cameras enable direct control of focus and aperture.

We are currently investigating several extensions to the basic approach. First, we are interested in further improving efficiency by taking advantage of the depth information from the camera's auto-focus sensors. Such information would let us save additional time, because we would only have to capture photos at focus settings that correspond to actual scene depths.

Second, we are generalizing the goal of light-efficient photography to reproduce arbitrary profiles of blur diameter *vs.* depth, rather than just reproducing the depth of field. For example, this method could be used to reproduce the defocus properties of the narrow-aperture photo entirely, including the slight defocus for background objects in Fig. 5.11.

# Chapter 6

# Time-Constrained Photography

> Time flies like an arrow. Fruit flies like a banana.
>
> Groucho Marx (1890–1977)

Suppose we have 100 ms to capture a given depth of field. What is the best way to capture a photo (or sequence of photos) to achieve this with highest possible signal-to-noise ratio (SNR)? In this chapter we generalize our previous light-efficient analysis (Chapter 5) to reconstruct the best in-focus photo given a fixed time budget. The key difference is that our restricted time budget in general prevents us from obtaining the desired exposure level for each photo, so we need also investigate the effect of manipulating exposure level. Manipulating exposure level leads to a tradeoff between noise and defocus, which we analyze by developing a detailed imaging model that predicts the expected reconstruction error of the in-focus image from any given sequence of photos. Our results suggest that unless the time budget is highly constrained (*e.g.*, below 1/30th of the time for the well-exposed time-optimal solution), the previous light-efficient sequence is optimal in these terms as well. For extreme cases, however, it is more beneficial to span the depth of field incompletely and accept some defocus in expectation.

## 6.1 Introduction

In the previous chapter we assumed that all photos were captured at an "optimal" exposure level of $L^*$, which means that every photo we considered was well-exposed and possessed good noise characteristics. Under this assumption, we showed that the time-optimal sequence spanning a particular DOF will generally involve multiple photos with large apertures (Fig. 6.1a). Since our analysis leads to the globally optimal solution, no other set of conventional photos at the same

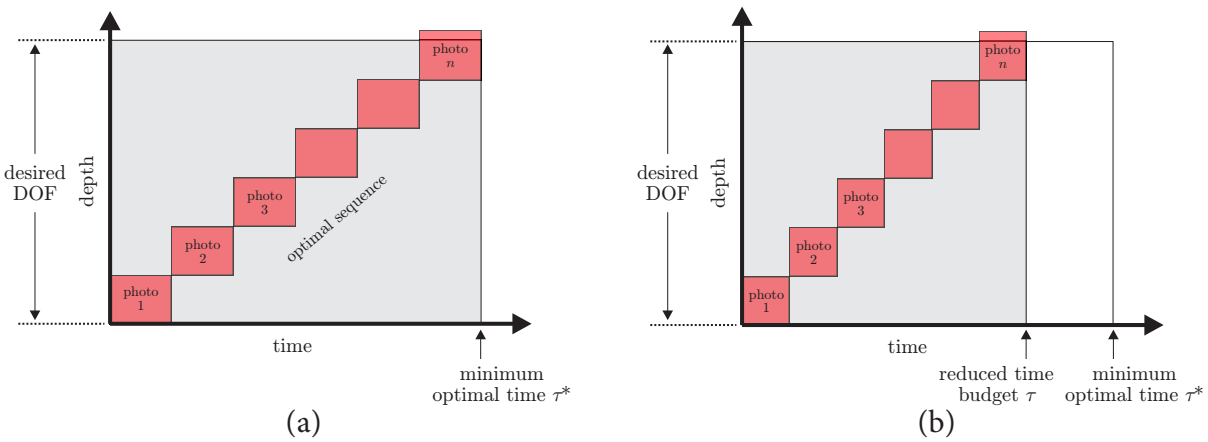



**Figure 6.1:** (a) Light-efficient photography, as a tiling of the DOF. The optimal sequence involves spanning the DOF using $n$ wide-aperture photos, each exposed at the desired level of $L^*$, and requires total time of $\tau^*$. As described in Sec. 5.4, the optimal sequence may slightly exceed the DOF. (b) Time-constrained photography. A simple strategy for meeting a reduced time budget is to reduce the exposure time of each photo proportionally. The tradeoff is that the reduced exposure level leads to increased noise.

exposure level can span the desired DOF faster than the total capture time of $\tau^*$ required by the light-efficient sequence.

Though applying our light-efficiency analysis can lead to greatly reduced total capture time compared to single-shot photography, what if we are constrained to even less time than the amount required by the optimal strategy—namely, what if $\tau < \tau^*$? This type of situation is common for poorly-illuminated moving subjects, where capturing a photo quickly enough to avoid motion blur means severely underexposing the subject. Since spanning the entire DOF and achieving well-exposed photos with an exposure level of $L^*$ requires total capture time of at least $\tau^*$, restricting ourselves to less capture time means sacrificing reconstruction quality in some sense.

To meet the more constrained time budget, the most obvious strategy is to reduce the exposure times of all photos in the light-efficient solution by a factor of $\tau/\tau^* < 1$ (Fig. 6.1b). Although these reduced-exposure photos will still span the DOF, they will be captured with a lower-than-optimal exposure level of $L = (\tau/\tau^*)L^*$, leading to increased noise.

A completely different strategy is to span the synthetic DOF incompletely, and expose the fewer remaining photos for longer, so that their noise level is reduced (Fig. 6.2). This strategy may at first seem counterintuitive, because it has the major disadvantage that the DOF is no longer fully spanned, and parts of the scene lying in the unspanned portion of the DOF will not be in-focus. Under the assumption the scene is distributed uniformly throughout the DOF, this strategy means accepting some level of defocus in expectation, because some of the scene will

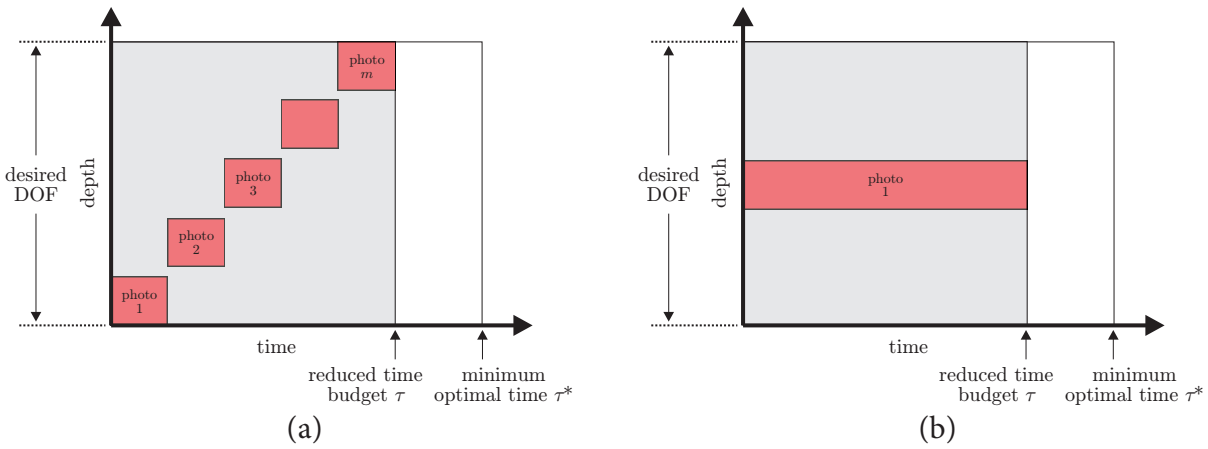



**Figure 6.2:** Alternative capture strategy for time-constrained photography. (a) By reducing the number of captured photos to $m < n$, each photo can be captured with higher exposure level and lower noise than the photos in Fig. 6.1b. This comes at the expense of not spanning the whole DOF, so parts of the scene will be defocused on average. (b) An extreme version of this strategy is to capture a single wide-aperture photo for the whole time interval. While its noise level will be significantly reduced compared to the photos in Fig. 6.1b and (a), only a small portion of the DOF will be spanned, therefore most of the scene will be defocused.

fall outside the synthetic DOF spanned by the sequence.

Under what conditions might it be valuable to not fully span the DOF? An illustrative example is the case where the time budget is so tightly restricted that all photos in the synthetic DOF are underexposed and consist only of quantization noise. By capturing just a single photo with increased exposure time instead (Fig. 6.2b), we have the potential to exceed the quantization noise and recover at least some low frequency signal, for an improvement in the overall reconstruction.

More generally, for a particular time budget, we study what capture sequence leads to the best synthetic-DOF reconstruction. Our analysis requires modeling the tradeoff between noise and defocus, each of which can be thought of as forms of image degradation affecting our ability to infer the ideal synthetic DOF photo. To quantify the benefit of a given capture sequence, we estimate the signal-to-noise ratio (SNR) of the reconstruction it implies, based on a detailed model we propose for the degradation process. More specifically, our analysis relies on explicit models for lens defocus and camera noise, and on a simplified model for the scene.

The closest related work is a recent report by Levin, *et al.* that compares several general families of camera designs and capture strategies [70], which makes use of a similar formulation as ours. A major difference in our approach is the way we model taking multiple photos: while we divide up the budget of capture time (at the expense of exposure level), they divide up their budget of sensor elements (at the expense of resolution). Moreover, while our main interest is

in evaluating how the optimal capture strategy for a conventional camera varies over different scenarios, they fix the capture scenario and determine the parameters for each of several camera designs in a one-time optimization.

Our work offers three main contributions. First, we develop a theory that generalizes light-efficient photography to situations where our time budget is constrained to less than the amount needed to capture the synthetic DOF at the desired exposure level. Second, we present a detailed framework for imaging that unifies the analysis of defocus and noise, that is simple enough to be evaluated analytically for a scene at given depth, but also leads to a practical reconstruction method. Third, our analysis shows that the near-optimal capture strategy can be expressed in very simple terms: we should use the wide aperture corresponding to our previous light-efficiency analysis, expose all photos equally, and space them evenly throughout the DOF. The only unspecified aspect of the strategy, how many wide-aperture photos to use, can be resolved by evaluating the SNR for each option according to our detailed imaging model.

## 6.2 Imaging Model with Defocus and Noise

To analyze the ability of a particular capture sequence to reconstruct the underlying ideal image, we develop a detailed image formation model, composed of explicit models for lens defocus and camera noise, and using a statistical model of the scene. Since we model each of these components as linear transformations or as Gaussian distributions, the overall model allows us to derive analytic expressions for the optimal reconstruction and for its SNR.

**Statistical scene model.** Our analysis assumes that the scene is a set of textured fronto-parallel planar patches located at random depths, distributed uniformly over the given DOF.[1] We specify the depth distribution in the corresponding space of focus settings (Sec. 5.2.1),

$$v \sim \text{Uniform}(\alpha, \beta) \ , \tag{6.1}$$

for a DOF spanning $[\alpha, \beta]$. Since we analyze a patch of the image at a time in practice, our scene model can be thought of as a fronto-parallel plane distributed randomly in depth.

We also assume that the scene texture follows natural image statistics [98]. While more general models are possible [70, 98], we define a prior distribution on the texture using a simple

[1] This implies either that the whole scene visible from the camera's viewpoint lies within the DOF, or that the boundaries of the DOF can be delineated in the images.

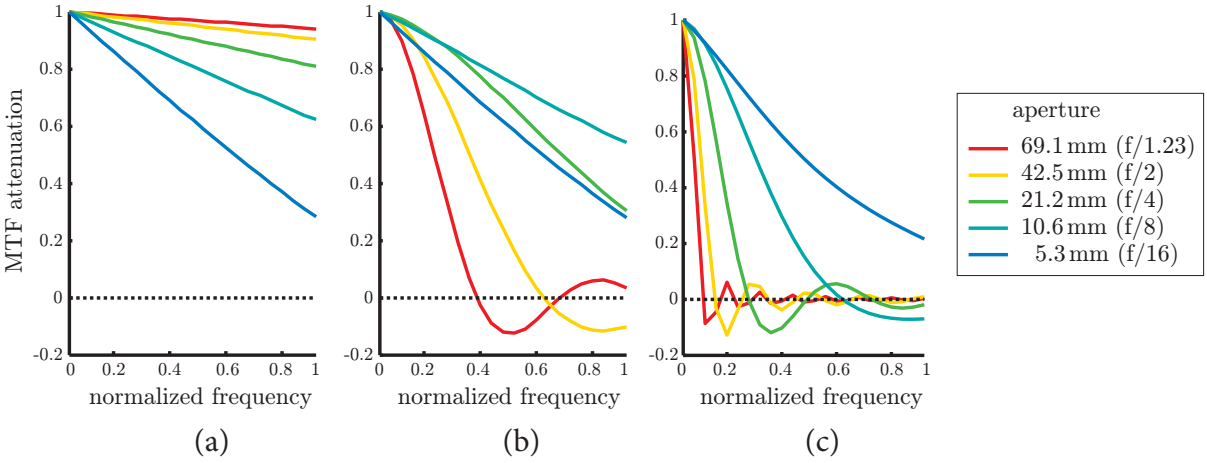



**Figure 6.3:** MTF for different apertures and levels of defocus, using the analytic model accounting for defocus and diffraction in [59]. We show the MTF for a planar scene 117 cm from the camera, with a focal length 85 mm. The DOF for an f/16 aperture focused at 117 cm is [110 cm, 124 cm]. The maximum frequency shown corresponds to a pixel spacing of 7.1 µm on the sensor. (a) Perfectly in-focus, limited only by diffraction. (b) Moderately defocused, 1 cm away from the in-focus plane. (c) More defocused, 5 cm away from the in-focus plane.

Gaussian for ease of analysis,

$$\boldsymbol{\beta}^* \sim \mathcal{N}\left(\boldsymbol{\mu}_{\text{prior}}, \mathbf{M}_{\text{prior}}\right) , \tag{6.2}$$

where $\boldsymbol{\beta}^*$ is the ideal in-focus photo. Following the formulation of Levin, *et al.* [69], we specify the parameters of this distribution in order to penalize the sum of squared image gradient magnitudes:

$$\boldsymbol{\mu}_{\text{prior}} = 0 \tag{6.3}$$

$$\mathbf{M}_{\text{prior}}^{-1} = \alpha(\mathbf{C}_x^T\mathbf{C}_x + \mathbf{C}_y^T\mathbf{C}_y) , \tag{6.4}$$

where $\mathbf{C}_x, \mathbf{C}_y$ are block-Toeplitz matrices representing convolution with the filter $[-1\ +1]$ over the $x, y$ dimensions of the image respectively. We determined the constant $\alpha = 0.004$ empirically, by matching $\mathbf{M}_{\text{prior}}$ to the variance of a large set of "natural" images generated randomly according to the standard $1/\nu$ frequency distribution [98].

**Defocused imaging using the lens MTF.** To model the effects of defocus on the scene, we characterize image formation in the frequency domain using the modulation transfer function (MTF). The MTF is defined as the signed magnitude of the Fourier transform of the 2D

point-spread function (Sec. 2.3.1), and can be interpreted as providing a set of per-frequency attenuation factors degrading the true frequency content of the ideal photo. We assume that the point-spread function is circularly symmetric, so we can express the MTF as a function of the radial frequency, $\nu = (\nu_x^2 + \nu_y^2)^{\frac{1}{2}}$.

Although many lens manufacturers provide MTF charts to characterize the performance of real lenses, these charts are not directly useful, because they only provide measurements for a few specific frequencies and assume an in-focus scene [7]. Instead, we used a classic analytic model for MTF developed by Hopkins [59], that accounts for the combined effects of defocus and diffraction. For a given blur diameter $\sigma$ and aperture diameter $D$, the expression for the MTF is:

$$\mathbf{mtf}(\nu) \;=\; \frac{4}{\pi^2 \sigma \nu} \int_0^{\sqrt{1-s(\nu)^2}} \sin\Big(\pi\sigma\nu\big[\sqrt{1-x^2} - s(\nu)\big]\Big)\,\mathrm{d}x \;\;, \tag{6.5}$$

where

$$s(\nu) \;=\; \nu \frac{F}{D}\left(\frac{\nu - F}{\nu}\right) \cdot 546\,\mathrm{nm} \tag{6.6}$$

is the normalized spatial frequency, expressed relative to the diffraction limit of the aperture. As before, $F$ is the focal length of the lens, and $\nu$ is the focus setting, defined by the distance from the sensor plane to the lens. We evaluated Eq. (6.5) using numerical integration (Fig. 6.3).

While we currently restrict our attention to an analytic lens MTF [59], nothing in our framework prevents us from using an empirically measured MTF, or an MTF that takes into account spatial variation over the image.[2]

**Linear image formation with noise.** By assuming that the scene is locally fronto-parallel, both the blur diameter and the MTF will remain constant for a given patch, for any image of the scene. In this simple case, we can express an observed photo $\mathbf{y}$ as a linear transformation of the ideal in-focus image $\boldsymbol{\beta}^*$, scaled by the relative exposure level, plus sensor noise:

$$\mathbf{y} \;=\; \underbrace{\left(\frac{L}{L^*}\right)}_{\substack{\text{relative}\\ \text{exposure}}} \cdot \underbrace{\mathcal{F}^{-1}\,\mathrm{diag}(\mathbf{mtf})\,\mathcal{F}}_{\substack{\text{composite}\\ \text{transform, } \mathbf{X}}} \cdot \underbrace{\boldsymbol{\beta}^*}_{\substack{\text{ideal}\\ \text{image}}} \;+\; \underbrace{\boldsymbol{\varepsilon}}_{\substack{\text{multiplicative and}\\ \text{additive noise}}} \;\;, \tag{6.7}$$

where $\mathcal{F}$ is the change-of-basis matrix corresponding to the discrete 2D Fourier transform, and $\mathrm{diag}(\cdot)$ creates a diagonal matrix from a vector. In effect, the composite transformation matrix

[2] Lens aberrations (Sec. 2.2.2) are strongest at the extremes of a large-aperture image, so real lenses actually have spatially-varying MTFs. This effect is especially pronounced for low-quality lenses, for which large apertures can have significantly inferior MTF performance even at the in-focus setting, contrary to the analytic model illustrated in Fig. 6.3a.

$\mathbf{X}$ performs three operations in sequence: (1) it transforms the photo to the Fourier domain, (2) it attenuates frequencies according to the MTF, and then (3) it returns it to the spatial domain.

We model the sensor noise using a zero-mean Gaussian, whose variance $\mathbf{M}_\mathbf{y}$ has both a signal-dependent multiplicative component, approximating Poisson-distributed shot noise, and a constant component, accounting for read noise and quantization [56, 76]:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}\Big(0\,,\, \underbrace{\sigma_s^2\,\mathrm{diag}(\mathbf{y}) + \sigma_c^2\,\mathbf{I}}_{\text{noise variance, } \mathbf{M}_\mathbf{y}}\Big)\ . \tag{6.8}$$

This two-parameter noise model accurately captures the noise characteristics of many real sensors [3, 76]. Note that to make the estimation of $\boldsymbol{\beta}^*$ tractable, we express multiplicative noise as a function of the noisy observed photo $\mathbf{y}$, rather than of the unknown noise-free observation $\left(\frac{L}{L^*}\right)\mathbf{X}\boldsymbol{\beta}^*$.

## 6.3 Reconstruction and SNR Analysis

In the context of our more detailed imaging model, synthetic DOF photography means estimating the underlying ideal image $\boldsymbol{\beta}^*$ from a set of $K$ input photos $\{\,\mathbf{y}_i\,\}$, each with different aperture diameters, exposure times, and focus settings, $\{\,\langle D_i\,,\,\frac{L_i}{{D_i}^2}\,,\,\nu_i\rangle\,\}$. Previously, we used compositing techniques [10] to associate a single input photo with each pixel in the result (Sec. 5.5). Here, however, our analysis goes beyond simple compositing, since it incorporates information from all input photos at every pixel.

Assuming that the scene depth is known, we can use Eq. (6.5) to compute the MTF for each input photo given its capture parameters. This gives rise to a system of $K$ linear models, $\{\,\mathbf{y}_i = \left(\frac{L_i}{L^*}\right)\mathbf{X}_i\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_i\,\}$, based on Eq. (6.7). Consequently, estimating the underlying ideal image $\boldsymbol{\beta}^*$ can be thought of as a problem closely related to depth-from-defocus methods based on restoration [42, 54, 62, 95, 107]. What these restoration methods have in common is the potential to both denoise and deconvolve the input.

**MAP reconstruction with known depth.** Because all quantities are linear and Gaussian, the *maximum a posteriori* (MAP) estimate for $\boldsymbol{\beta}^*$ can be computed using a straightforward analytic formula, and will itself follow a Gaussian distribution [112]:

$$\hat{\boldsymbol{\beta}}_{\mathrm{MAP}} \sim \mathcal{N}\Big(\ \mathbf{M}_{\hat{\boldsymbol{\beta}}}\cdot\big[\mathbf{M}_{\mathrm{prior}}^{-1}\boldsymbol{\mu}_{\mathrm{prior}} + \textstyle\sum_{i=1}^{K}\mathbf{X}_i^T\mathbf{W}_i\mathbf{y}_i\big]\ ,\ \mathbf{M}_{\hat{\boldsymbol{\beta}}}\ \Big)\ , \tag{6.9}$$

where

$$\mathbf{M}_{\hat{\boldsymbol{\beta}}} = \left[\mathbf{M}_{\text{prior}}^{-1} + \sum_{i=1}^{K} \mathbf{X}_i^T \mathbf{W}_i \mathbf{X}_i\right]^{-1} \tag{6.10}$$

and $\mathbf{W}_i = \mathbf{M}_{\mathbf{y}_i}^{-1}$ are weights set according to the inverse variance of image noise. This computation amounts to a solving weighted linear least-squares problem, regularized by our natural image prior. Note that a completely "uninformative" prior corresponds to parameters $\boldsymbol{\mu}_{\text{prior}} = 0$ and $\mathbf{M}_{\text{prior}}^{-1} = 0$.

**SNR analysis.** To determine the SNR of the optimal reconstruction $\hat{\boldsymbol{\beta}}_{\text{MAP}}$, we can derive a simple formula based on its analytically computed variance $\mathbf{M}_{\hat{\boldsymbol{\beta}}}$,

$$\text{SNR} = \frac{\|\hat{\boldsymbol{\beta}}_{\text{MAP}}\|^2}{\text{E}\left[\|\hat{\boldsymbol{\beta}}_{\text{MAP}} - \boldsymbol{\beta}^*\|^2\right]} = \frac{\|\hat{\boldsymbol{\beta}}_{\text{MAP}}\|^2}{\text{tr}(\mathbf{M}_{\hat{\boldsymbol{\beta}}})}\ , \tag{6.11}$$

where $\text{tr}(\cdot)$ is the trace operator that sums along the diagonal.

**SNR analysis over depth.** So far, we have shown how to derive the optimal reconstruction and SNR for a particular capture sequence, for a fronto-parallel scene at a known depth. However, if the scene depth were truly known in advance, the optimal strategy would be the trivial approach of focusing at the known scene depth, then capturing a single wide-aperture photo for the entire time budget.

Instead, our main interest is characterizing capture sequences whose reconstruction properties are good in *expectation* over the random distribution of all scenes throughout the DOF (Eq. (6.1)). Since this expectation has no closed form, we evaluate it numerically, using Monte Carlo integration over depth. From Eq. (6.11), the expected SNR can be computed as

$$\text{E}\left[\text{SNR}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\|\hat{\boldsymbol{\beta}}_{\text{MAP}}(v_i)\|^2}{\text{tr}(\mathbf{M}_{\hat{\boldsymbol{\beta}}}(v_i))}\ , \tag{6.12}$$

where the depth samples $\{v_i\}$ follow the uniform random distribution from Eq. (6.1). In practice we compute Eq. (6.12) by sampling the DOF uniformly at 1000 different depths.

## 6.4 Candidate Sequences for Reconstruction

To help us analyze what capture sequences are optimal for time-constrained photography, we make several mild assumptions that greatly restrict the space of sequences we need to consider:

- **Equal apertures** First, we limit our analysis to continuously-variable apertures, because the optimal sequences resulting from Theorem 1 (p. 111) have the especially simple property that all photos use the same aperture.

  As a consequence, any time-constrained sequence that is optimal according to SNR must also consist of photos with equal apertures. This is because, according to Theorem 1, for any sequence not having equal apertures, there must be some equal-aperture sequence spanning an equivalent DOF in less total time, at the same exposure level. In the context of our fixed time budget, this equal aperture sequence corresponds to additional exposure time per photo, leading to increased exposure level and higher SNR.[3]

- **Equal exposure levels** We assume all photos in the optimal time-constrained sequence are captured with the same exposure level. Because our scene model is distributed over the DOF uniformly at random, we have no reason to collect more photons corresponding to a particular subinterval of the DOF at the expense of other areas. In expectation over our random scene model, SNR will therefore be highest when all photos are captured with equal exposure levels.

- **Even focus distribution** For the same reason, because our random scene model gives us no reason to concentrate on any particular DOF sub-interval, we assume that the focus settings for the optimal time-constrained sequence are distributed evenly throughout the DOF. More specifically, for a sequence consisting of $n$ photos, we choose focus settings corresponding to the narrow aperture photos that would be required to evenly span the DOF without gaps (Eq. (D.1)).

Based on these assumptions, all candidates for the optimal time-constrained capture sequence can be described using two parameters: (1) the aperture, and (2) the number of equal-aperture photos in the input sequence. Given these parameters, our simple assumptions fully define the exposure times and focus settings required for all photos in the input sequence.

Although our assumptions greatly reduce the space of capture sequences under consideration, the parameters that remain still provide a rich range of strategies exploring the most important tradeoff of time-constrained photography—balancing exposure level and DOF coverage. By manipulating these parameters, we are selecting between the two extremes of capturing many noisy photos spanning the whole DOF (Fig. 6.1b) and capturing a single photo with lower noise, but increased defocus (Fig. 6.2b).

[3] This argument relies on the assumption that the SNRs of the reconstructions for two sequences with "equivalent" sized DOFs are comparable, despite the fact that these DOFs may be non-contiguous. This simplification is reasonable, given the fact that our scene model is randomly distributed throughout the DOF, and we evenly distribute the focus settings of the sequence as well.

## 6.5 Results and Discussion

To investigate capture strategies for time-constrained photography, we modeled a particular DOF, camera, and underlying scene in simulation. We took care that all the simulation parameters correspond to realistic cameras and capture configurations used for synthetic DOF photography (Sec. 5.6).

More specifically, we used the same DOF and camera parameters as in Figs. 5.4–5.5, which themselves were inspired by the HAMSTER and FACE datasets (Sec. 5.6). We set the DOF to be $[110\,\text{cm}, 124\,\text{cm}]$ and we model a continuous-aperture lens with a focal length of 85 mm and a maximum aperture of f/1.2. As described in Fig. 5.4, the time-optimal sequence for this configuration consists of 13 photos using a wide aperture of f/1.23 (69.1 mm), for a total capture time of $\tau^* = 115\,\text{ms}$.

To model the camera sensor, we assumed that the pixels are spaced 7.1 μm apart, corresponding to a 16 MP sensor that is 35.4 mm wide. We determined parameters for our noise model in Eq. (6.8) according to an empirical evaluation of the CMOS sensor for the Canon EOS-1D Mark II [3]. In particular, we used the parameters $\sigma_s = 0.049$ and $\sigma_c = 0.073$, where pixel values are scaled to the range $[0, 255]$.

For the underlying scene, we used a 45-pixel 1D image patch that we generated randomly according to our natural image prior (Sec. 6.2). In practice, the relative SNRs of different capture strategies are insensitive to the underlying scene, because the only scene-specific feature of our analysis is the between-pixel variation in shot noise, whose influence on SNR is limited. We also found that for the purpose of comparing different time-constrained capture strategies, neither using larger 1D patches nor using 2D patches produced any significant qualitative difference in overall results.

**Largest aperture yields highest SNR.** For our first experiment, we explored the influence of aperture diameter on SNR. We exhaustively tested all time-constrained capture strategies described in Sec. 6.4, varying the available parameters of aperture and number of photos, for a several different constrained time budgets (Figs. 6.4 and 6.5). For every time budget we tested, we found that the optimal SNR increased monotonically with aperture size. More specifically, the highest-SNR capture sequence for a given aperture exceeded the SNR of all sequences captured using narrower apertures.

Intuitively, the reason why wider apertures lead to higher SNR is similar to our argument for the optimality of equal apertures (Sec. 6.4). Recall that according to Theorem 1, the time-optimal

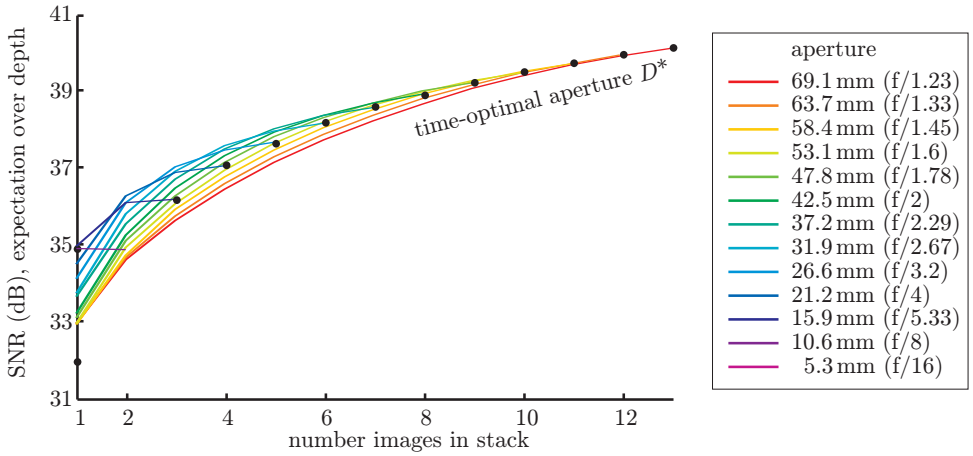



**Figure 6.4:** SNR for the two-parameter set of capture strategies that we consider, for a reduced time budget of $\tau = \frac{1}{3}\tau^*$. Each series corresponds to a different aperture, and defines a set of equal-aperture capture sequences, ranging from a single photo to fully spanning the DOF. The black points indicate the optimal number of photos to use for each aperture. The highest overall SNR is achieved by fully spanning the DOF using 13 photos at the widest aperture.

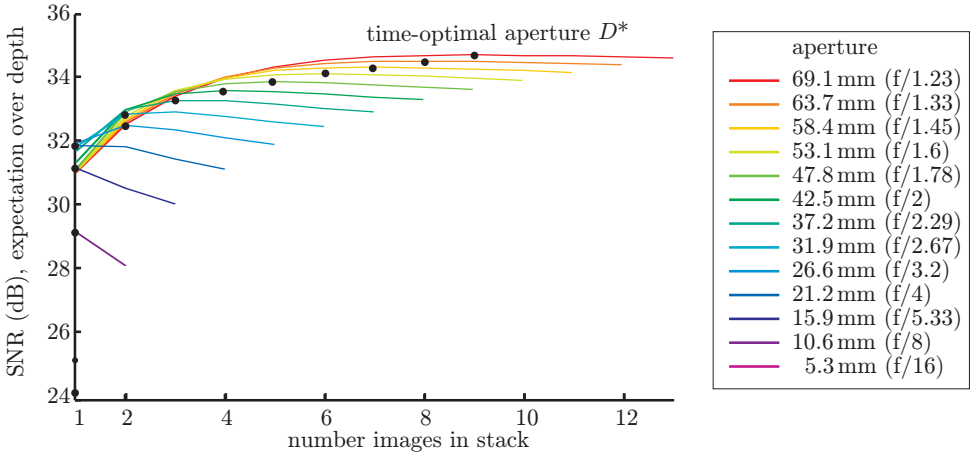



**Figure 6.5:** SNR for the two-parameter set of capture strategies that we consider, for a reduced time budget of $\tau = \frac{1}{30}\tau^*$. Each series corresponds to a different aperture, and defines a set of equal-aperture capture sequences, ranging from a single photo to fully spanning the DOF. The black points indicate the optimal number of photos to use for each aperture. The highest overall SNR is achieved by *incompletely* spanning the DOF using 9 photos at the widest aperture.

sequence consists of wide-aperture photos whose apertures have diameter $D^*$. Consequently, any capture sequence consisting of narrower-aperture photos must be sub-optimal, in the sense that there is some wider-aperture sequence that can span an equivalent DOF in less total time, at the same exposure level. Therefore, under our fixed time budget, the time savings realized from wider apertures can be traded for higher exposure level and therefore higher SNR.

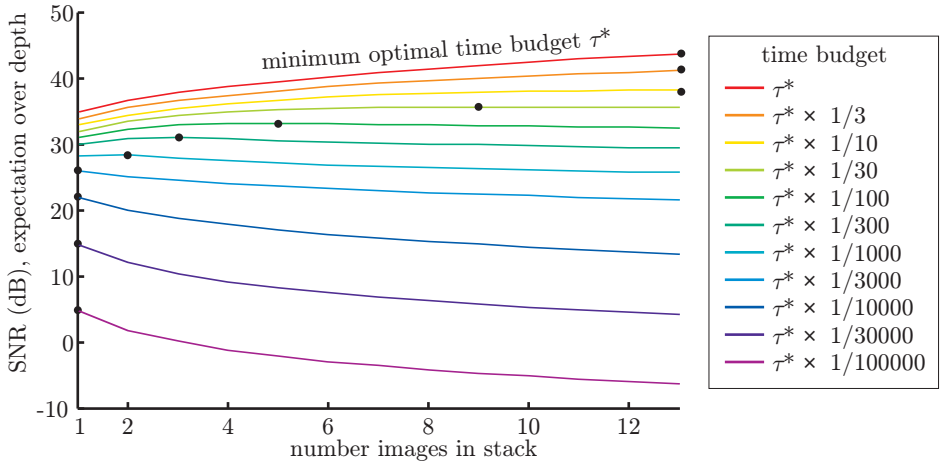



**Figure 6.6:** SNR for different numbers of photos in the capture sequence. Each series corresponds to a different time budget, and illustrates how the SNR of the reconstruction changes over different numbers of photos in the capture sequence. The black points indicate the optimal number of photos for each time budget, each of which is taken with an f/1.23 aperture.

**Densest focal stacks yield highest SNR (unless time extremely limited).** For our second experiment, we investigated how the optimal SNR capture strategy varies over different time budgets. Following the previous discussion, we fixed the aperture diameter to $D^*$ = f/1.23, corresponding to the time-optimal sequence from Theorem 1. For each time budget, we varied the number of photos taken, which affects both the exposure level and the DOF coverage, and defines a range of capture strategies (Fig. 6.6).

In the case where the time budget is not restricted, meaning that the full time of $\tau^*$ is available, our results show that the sequence with the highest SNR fully spans the DOF using 13 photos, reproducing our previous result from Theorem 1. In other words, when there is enough time to capture all photos at their optimal exposure level, the time-optimal capture sequence from our light-efficiency analysis is also the sequence with the highest SNR.

For this example, fully spanning the DOF remains the highest-SNR capture strategy until the time budget drops significantly. Only at a time budget of $\frac{1}{30}\tau^*$ or below does the relative noise level become high enough that incompletely spanning the DOF to reduce per-photo noise leads to an overall improvement in SNR. In the limit, for a time budget of $\frac{1}{3000}\tau^*$ or below, the signal is so degraded that little more than the DC component can be recovered, and the capture sequence with highest SNR consists of single photo exposed for the entire time budget.

**Effect of the natural image prior.** To evaluate the effect of our natural image prior, we redid the previous experiment without any prior image model (Fig. 6.7). As our results show, in

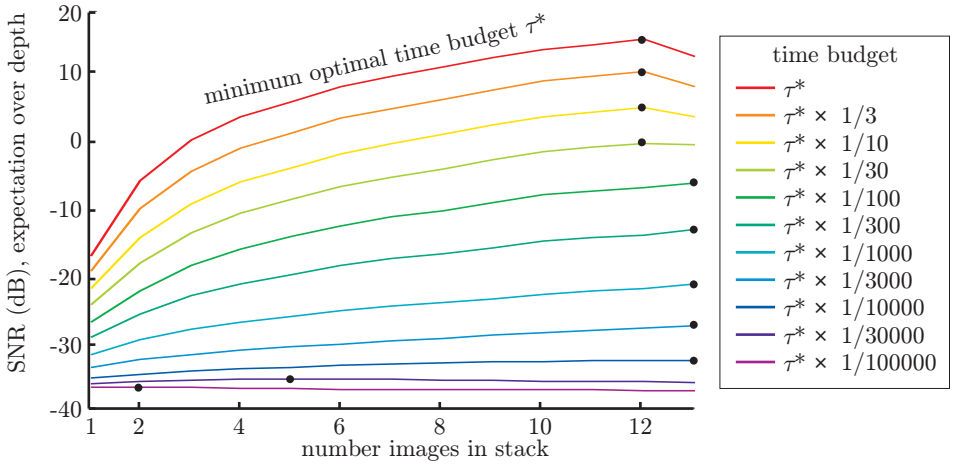



**Figure 6.7:** Effect of removing the natural image prior. We evaluated the SNR of the same capture sequences as in Fig. 6.6, but without an image prior. Each series corresponds to a different time budget, and the black points indicate the optimal number of photos for each time budget, each of which is taken with an f/1.23 aperture.

the absence of the prior, the relative variation of SNR across different numbers of photos in the capture sequence depends far less on the time budget. In particular, for all but the most extreme time budgets, the capture sequence with the highest SNR involves 12 or 13 wide-aperture photos, spanning the DOF (or nearly so).[4]

Intuitively, whenever we use a smaller number of photos and incompletely span the DOF, the loss of high-frequency information caused by the expected defocus will lead to large reconstruction error in those high frequencies. Using an image prior helps us discount this error by assigning low likelihood to high frequency image content. However, without such a prior, the reconstruction of defocused high frequencies will significantly magnify image noise and reduce the overall SNR. Therefore, without an image prior, much greater coverage of the DOF is required.

**Effect of overhead.** For our final experiment, we simulated the effect of overhead for a 60 fps camera (16.7 ms overhead per photo), but otherwise for the same scene and DOF as before (Fig. 6.8). With this level of overhead, our previous light-efficiency analysis tells us that the time-optimal sequence involves 10 photos, each with a narrower aperture of f/1.6, for a total capture time of $\tau^* = 300$ ms (Fig. 5.5).

As shown in Fig. 6.8, when the time budget is significantly reduced, many multi-photo cap-

[4]The fact that the highest-SNR sequence does not fully span the DOF as in Fig. 6.6 is another "discretization" artifact due to the specific form of our MTF. The oscillating tails of the MTF interact with the spacing of the focus settings in the input, which can potentially lead to higher SNR for slightly reduced DOF coverage.

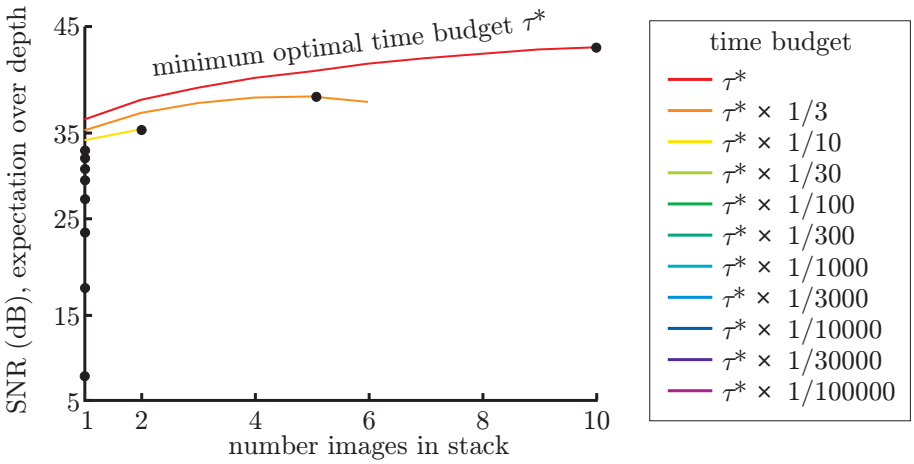



**Figure 6.8:** Effect of camera overhead. We evaluated SNR for the same scene and DOF as in Fig. 6.6, but simulating the effect of 60 fps overhead (16.7 ms per photo). The overhead affects the time-optimal sequence (Fig. 5.5), which in turn reduces the aperture diameter to f/1.6. The overhead also affects the relative exposure level for each input photo, which reduces the SNR for sequences with more photos. Each series corresponds to a different time budget, and the black points indicate the optimal number of photos for each time budget, each of which is taken with an f/1.6 aperture.

ture strategies simply cannot be realized, because overhead alone exceeds the time budget available. For this example, reducing the time budget below $\frac{1}{10}\tau^*$ means that we only have time to capture a single photo.

When overhead is considered, capture sequences consisting of fewer photos will enjoy relatively higher exposure levels, because they spend a greater fraction of the total capture time actually collecting light. This leads to an effective increase in SNR for input sequences consisting of fewer input photos.

In general, the results with camera overhead are qualitatively similar to the wider-aperture overhead-free case shown in Fig. 6.6. As before, for an unconstrained time budget of $\tau^*$, the optimal-SNR sequence reproduces the previous time-optimal result. Furthermore, as the time budget is reduced, the optimal strategy also shifts toward capturing single photo exposed for the entire time budget.

## Summary

Although our results are still preliminary, and we have only tested our analysis in simulation, we believe that our approach for time-constrained photography provides the groundwork necessary for a more general analysis of conventional photography. This new analysis accounts for the tradeoffs between noise, defocus, and exposure level in a unified framework. As such, it subsumes our previous geometric view of DOF and exposure (Chapter 5) according to a more

detailed model based on the MTF and an explicit model for noise.

So far, in analyzing the relative SNR of different capture strategies, we have assumed that we are able to accurately estimate the depth of the scene. Our ongoing experiments suggest that standard depth-from-defocus methods (*e.g.*, [120]) may actually be adequate for this purpose. Note that when such methods are ambiguous (*e.g.*, due to lack of texture, or high noise), accurate depth estimation will have less effect on the quality of the reconstruction.

For improved performance at capture time, however, it could be valuable to integrate depth recovery into the MAP reconstruction directly. In particular, by explicitly modeling the depth uncertainty of the input photos, we should be able to derive a reconstruction whose SNR is optimal over the depth distribution of the scene, conditioned on our observations.

We are also interested in addressing practical reconstruction challenges such generalizing our reconstruction method to handle image patches that are not fronto-parallel (Sec. 2.3) or contain depth discontinuities. In general, when the scene is significantly tilted or contains depth discontinuities, our frequency-based analysis, which locally assumes spatially-invariant convolution will break down. From the point of view of generating visually realistic images it may be valuable to fit a more detailed depth model to each image patch [71], or to revisit gradient-based techniques for blending the reconstruction results from overlapping image patches [10].

# Chapter 7

# Conclusions

> Seeing much, suffering much, and studying much, are the three pillars of learning.
>
> Benjamin Disraeli (1804–1881)

> The supreme accomplishment is to blur the line between work and play.
>
> Arnold Toynbee (1889–1975)

It is an exciting time to be thinking about photography. The computational approach that has coalesced in the past fifteen years has prompted researchers across disciplines to revisit previous assumptions about what can be accomplished with cameras, and what limits in photography are truly fundamental.

Some of the techniques born from this programme, such high dynamic range photography [31, 78] and automatic image stitching for panoramas [26], have already been adopted by a worldwide cadre of technically-minded photographers. It is not difficult to imagine these methods becoming more widespread, nor that our current approach to photography may be poised for an even more radical shift.

The thrust of this dissertation has been an extended argument for replacing one-shot photography with multiple photographs captured with varying camera settings. As our methods demonstrate, by applying computation and analyzing the defocus characteristics of sequences of photos, we can realize richer capabilities that go beyond what is possible with conventional photography. In particular, we have shown that:

- We can recover detailed 3D structure at the level of individual pixels, even for scenes with complex geometry (Chapter 3).
- We can give the photographer the flexibility to manipulate all camera controls in post-processing, by using the standard "aperture bracketing" function (Chapter 4).

- We can capture a well-exposed photo with a given depth of field faster than conventional photography allows, by taking multiple large-aperture photos (Chapter 5).
- The sequence of photos best-suited for capturing a given depth of field in a specified time budget is the densest possible focal stack, up to the limit imposed by overhead and additive noise (Chapter 6).

While all of our approaches can be applied immediately, using existing digital cameras without modification, we stand to benefit from lower-overhead control of the camera settings and from increased capture speed.

New computational techniques for photography have generally been inspired by real or anticipated advances in camera technology. For example, the current glut of sensor resolution has given rise to methods trading resolution for measurements along other dimensions, such as viewing direction [9, 47, 73, 75, 85, 115]. We believe that a coming excess of *capture speed* will inspire new methods capturing multiple high-resolution photographs with varying camera settings, along the lines of the methods we suggest. The trend toward higher capture speed is supported by the recent Casio EX-F1 digital camera, which can record 6 MP photos at 60 fps, and by high-definition digital video cameras such as the Red One, which blur the line between video and high-speed photography.

## 7.1 Future Research Directions

In the years ahead, we are interested in exploring new aspects of variable-aperture photography, as well as more general capture scenarios, including changes in viewpoint and the addition of controllable illumination. The framework of enhancing photography with computation encompasses a vast space of possible camera designs, and there are a number of outstanding problems relating to the nature of this space.

**Efficient systems for computational photography.** Because many scenes are inherently dynamic, including scenes with people or flowing water, the capture time available is often limited by the amount of motion blur we can tolerate. A fundamental open question is how to make best use of the capture time and sensor resolution available, given a particular camera design. This dissertation has made several first steps in this direction, showing that when the camera is programmable, it can be significantly more efficient to divide the available time and capture multiple photos with different settings (Chapters 5–6). Building on this work, we would like to devise more general schemes for improving efficiency. Another interesting extension would

be to devise adaptive capture methods, where an online analysis of the photos captured so far would determine the next camera setting in light of the remaining capture time.

Beyond these first problems, which all assume a given camera model, we would also like to investigate the optimal spatial configurations of the sensor elements themselves. One would expect such analysis to vary according to the goals we set, for example, accurately reconstructing 3D depth or capturing a given in-focus region. This work is also exciting for its potential to provide a unified, fair comparison among a disparate set of existing imaging systems.

**Focal stack stereo.** A fundamental limitation of conventional photography is that the accuracy of 3D reconstruction from a given viewpoint depends on the size of the lens aperture [99]. Since large lenses are expensive to build, a more scalable idea is to combine focal stacks from different viewpoints instead, treating each photo as a discrete sample of some larger virtual aperture. The same idea applies to microscopy applications [75], but because microscopy images are orthographic, having multiple viewpoints here corresponds to having multiple *angles* of view.

While viewpoint and defocus provide complementary information, existing methods have analyzed this combination in a limited way [13]. By taking a unified view and revisiting the underlying geometry, we hope to develop more sophisticated 3D reconstruction methods that incorporate the advantages of defocus-based analysis, but which also provide much greater potential depth resolution. One application that could benefit especially from this work is stereo microscopy in medical imaging.

**Global shape from focus and defocus.** Although reconstruction methods such as confocal stereo (Chapter 3) provide strong depth cues at the level of individual pixels, we would ideally be able to reconstruct 3D scenes according to a model of defocus that fully accounts for occlusion and global scene geometry [16]. Preliminary steps have already been made in this direction [16, 42], but the proposed methods are computationally intensive, and do not seem to scale to higher resolutions or to scenes composed of more than a few surfaces. One approach that seems to hold promise is a space-sweep approach analogous to voxel-based stereo [68].

**New visual effects.** When capturing a scene over a variety of camera settings, we obtain multi-dimensional datasets that are much richer than standard photos, since every pixel records the integration of a different cone of rays. Although we have already described methods for compositing and resynthesis from the more compact scene descriptions we recover (Chapters 4–5), it could also be valuable to interpolate the raw pixel data more directly, analogous to light field

rendering techniques [48, 74].

It would also be interesting to use the captured dataset to develop new visualizations of the scene, or to develop tools would enable photographers to create new artistic effects. For example, a photographer might wish to specify a non-physical depth of field that follows a curved surface in the scene, but have defocus respect the distorted optics as closely as possible.

# Appendix A

# Evaluation of Relative Exitance Recovery

To obtain a more quantitative evaluation of how well the relative exitance $\mathbf{R}_{xy}(\alpha, f)$ can be recovered, and to validate that it not sensitive to experimental conditions, we ran several additional experiments.

**Experiment 1: Repeatability.** To test repeatability across different captures under fluorescent lighting, we repeated 5 trials of the radiometric calibration described in Sec. 3.5 for a diffuse white plane, for 13 aperture settings at a fixed focus setting. We used a Canon EF85mm 1.2L lens, as in Sec. 3.8.

For each pixel and aperture setting, we measured the standard deviation of $R$ over the 5 trials, as a fraction of the mean. Over all pixels, the median of this fraction was 0.51 % and its RMS measure is 0.59 %. This indicates good repeatability after correcting for lighting fluctuations.

**Experiment 2: Stability over new scenes.** To validate that the ratio $R$ measured in radiometric calibration can be applied to new scenes, we redid the previous calibration for three additional scenes, all at the same focus setting. To create the new scenes, we used the same diffuse white calibration plane, but tilted it (about 45°) to different 3D configurations, yielding calibration images with different shading.

For each of the three new scenes, we computed the relative errors between the measured ratios $R$, and the corresponding ratios from the previous calibration (Experiment 1, trial 1). The aggregate results (Table A.1) show that the median magnitude of the relative error is 1–2 gray levels out of 255.

**Experiment 3: Legacy calibration.** We also compared the radiometric calibration from these experiments to the calibration used in Sec. 3.8, captured several years beforehand using

**Table A.1:** Evaluating the stability of the radiometric calibration over new scenes.

| new scene | median abs. relative error | RMS relative error |
|---|---|---|
| tilted plane #1 | 0.76 % | 1.29 % |
| tilted plane #2 | 0.83 % | 1.63 % |
| tilted plane #3 | 0.78 % | 1.28 % |

the same lens.

For each pixel and aperture setting, we computed the relative error between the ratio $R$ as originally computed, and the corresponding ratio from the radiometric calibration in Experiment 1, trial 1. Over all pixels, the median magnitude of the relative error was 1.10 % and its RMS measure is 2.21 %. This agreement is good, given the fact that we did not use the same focus setting or calibration target for this experiment.

**Experiment 4: Different lens.** As a final test, we redid the calibration in Experiment 1 using a different lens, but of the same model, with the calibration target placed at approximately the same distance.

We again computed relative errors between the recovered ratios $R$, and the corresponding ratios from the Experiment 1, trial 1. Over the entire image, the median magnitude of the relative error was 0.87 % and its RMS measure is 1.78 %. This error level is on the same order as Experiment 2, suggesting that calibration parameters persist across lenses, and that radiometric calibration can be done just once for each *model* of lens, provided that manufacturing quality is high.

# Appendix B

# Conditions for Equi-Blur Constancy

In Sec. 3.7.2 we described how it is possible to approximate a pixel's AFI using a set of equi-blur regions where color and intensity remain constant. Here we establish conditions **C1**–**C5** under which this approximation becomes exact.

Suppose that scene point $\hat{\mathbf{p}}$ is in perfect focus for setting $\hat{f}$ and projects to point $(x, y)$ on the sensor plane. Now suppose we defocus the lens to some setting $(\alpha, f)$ (Fig. B.1). We assume the following condition:

**C1.** Lens defocus can be described using the thin-lens model [16, 105].

Then the image irradiance at $(x, y)$ is

$$E_{\alpha f}(x, y) \;=\; \frac{\pi\left(\frac{F}{2\alpha}\right)^2 \cos^3\theta}{z^2} \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} \frac{L(\mathbf{q}(\omega),\, \omega)\;\cos\beta(\omega)}{\|\, \mathcal{C}_{xy}(\alpha, f)\,\|}\, \mathrm{d}\omega \;, \tag{B.1}$$

where $\frac{F}{\alpha}$ is the aperture diameter; $\theta$ is the angle between the optical axis and the ray connecting $(x, y)$ and the lens center, $C$; $z = (\frac{1}{\lambda} - \frac{1}{\mathrm{dist}(f)})^{-1}$ is the distance from the aperture to the sensor plane; $\mathcal{C}_{xy}(\alpha, f)$ is the cone converging to the in-focus scene point $\mathbf{p}$, which lies off the scene surface; $\mathbf{q}(\omega)$ is the intersection of the scene with the ray from $\mathbf{p}$ in direction $\omega$; $L(\mathbf{q}(\omega), \omega)$ is the outgoing radiance from $\mathbf{q}(\omega)$ in direction $\omega$; and $\beta(\omega)$ is the angle between the optical axis and the ray connecting $\mathbf{p}$ to $\mathbf{q}(\omega)$.

Our goal is to show that $E_{\alpha f}(x, y)$ in Eq. (B.1) is constant for all points in an equi-blur region. That is, if $(\alpha', f')$ is also in the same equi-blur region as $(\alpha, f)$, with

$$b_{\alpha' f'} \;=\; b_{\alpha f} \;=\; \frac{F}{\alpha}\, \frac{|\mathrm{dist}(\hat{f}) - \mathrm{dist}(f)|}{\mathrm{dist}(f)} \;, \tag{B.2}$$

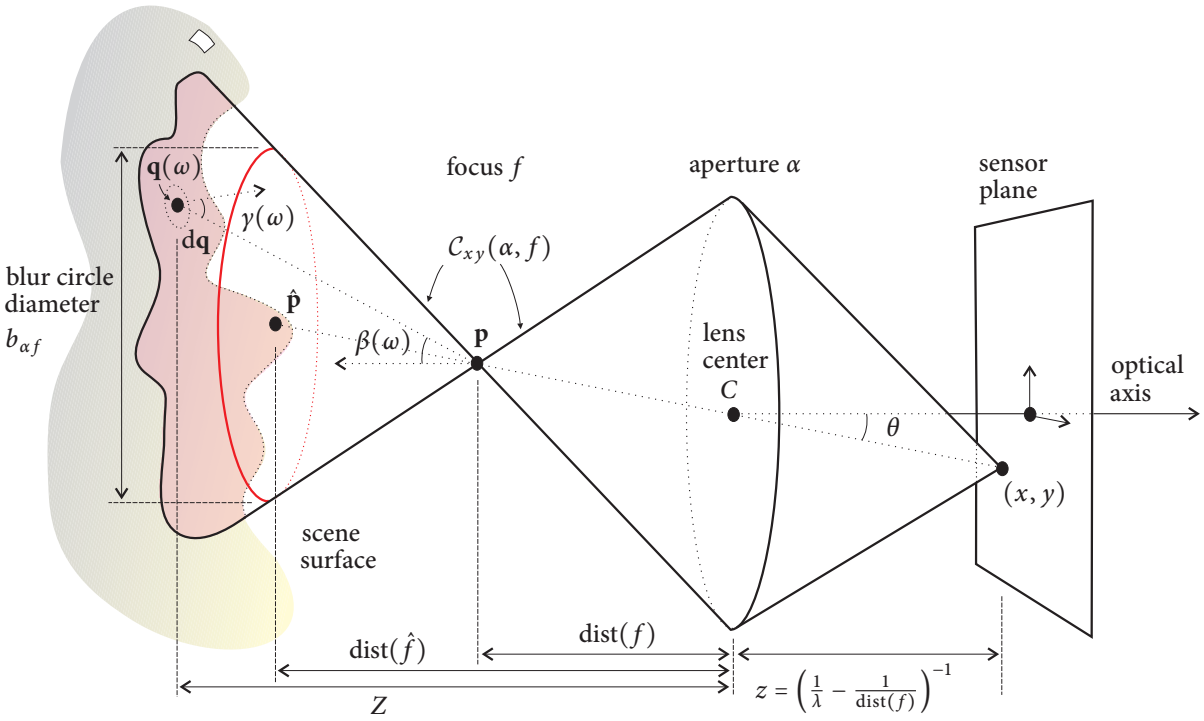



**Figure B.1:** Thin lens imaging model for defocus [16, 105]. At an out-of-focus setting $f$, a point on the sensor plane $(x, y)$ integrates radiance from a region of the scene as shown. By contrast, at the perfect focus setting $\hat{f}$, all irradiance at $(x, y)$ would be due to scene point $\hat{\mathbf{p}}$. We characterize the level of "blur" using a fronto-parallel circle with diameter $b_{\alpha f}$ and centered on $\hat{\mathbf{p}}$, which approximates the intersection of cone $\mathcal{C}_{xy}(\alpha, f)$ with the scene surface. In our approximate model, the irradiance integrated at $(x, y)$ will remain constant for any other lens setting $(\alpha', f')$ yielding the same blur circle diameter.

then $E_{\alpha' f'}(x, y) = E_{\alpha f}(x, y)$. We show this by showing that $E_{\alpha f}(x, y)$ is independent of $(\alpha, f)$, for all $(\alpha, f)$ in the same equi-blur region.

To do this, first we assume the following condition:

> **C2.** From any scene point, the solid angle subtended by the largest aperture approaches zero, *i.e.*, $\|\mathcal{C}_{xy}(\alpha, f)\| \to 0$.

This allow us to simplify Eq. (B.1), because it implies that $\beta(\omega) \to \theta$, giving

$$E_{\alpha f}(x, y) = \frac{\pi\left(\frac{F}{2\alpha}\right)^2 \cos^4\theta}{z^2} \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} \frac{L(\mathbf{q}(\omega), \omega)}{\|\mathcal{C}_{xy}(\alpha, f)\|} \, \mathrm{d}\omega \ . \tag{B.3}$$

Note that the factor outside the integral in Eq. (B.3) is independent of the scene and accounted for by radiometric calibration (Sec. 3.5). Therefore this factor is independent of $(\alpha, f)$ and it suffices to show that the integral is independent of $(\alpha, f)$ in the equi-blur region.

The integrand in Eq. (B.3) is simply the contribution to irradiance of a differential patch d$\mathbf{q}$,

centered on point $\mathbf{q}(\omega)$ and subtending a solid angle of $\mathrm{d}\omega$ from $\mathbf{p}$. Now consider the following two conditions:

**C3.** The outgoing radiance for any defocused scene point is constant within the cone subtended by the largest aperture, *i.e.*, $L\left(\mathbf{q}(\omega), \omega\right) = L\left(\mathbf{q}(\omega)\right)$.

**C4.** For any defocused scene point, the cone subtended by the largest aperture does not intersect the scene elsewhere.

Note that conditions **C3**–**C4** are the same conditions required by confocal constancy (Sec. 3.3), but applied to all points in the defocused region of the scene. The radiance of the differential patch, namely the factor $L(\mathbf{q}(\omega), \omega)$ in Eq. (B.3), is independent of $(\alpha, f)$. Hence it suffices to show that the geometric factor $\frac{\mathrm{d}\omega}{\| \mathcal{C}_{xy}(\alpha,f) \|}$ is independent of $(\alpha, f)$ in the same equi-blur region.

From the definition of solid angle, this factor is given by

$$\frac{\mathrm{d}\omega}{\| \mathcal{C}_{xy}(\alpha, f) \|} = \frac{\mathrm{d}\mathbf{q} \cos\gamma(\omega) \cos^2\beta(\omega)}{(Z - \mathrm{dist}(f))^2} \cdot \frac{\mathrm{dist}(f)^2}{\pi\left(\frac{F}{2\alpha}\right)^2 \cos^3\theta} \ , \tag{B.4}$$

where $\mathrm{dist}(f)$ is the distance from $\mathbf{p}$ to the aperture; $Z$ is the distance from $\mathbf{q}(\omega)$ to the aperture; and $\gamma(\omega)$ is the angle between the surface normal of $\mathrm{d}\mathbf{q}$ and the ray connecting $\mathbf{q}(\omega)$ to $\mathbf{p}$.

Now assume that the following condition also holds:

**C5.** Depth variations for points within the defocused region of the scene approach zero, *i.e.*, $Z \rightarrow \mathrm{dist}(\hat{f})$.

This condition implies that the depth, $Z$, of the differential patch $\mathrm{d}\mathbf{q}$ can be approximated by the distance to the scene point $\hat{\mathbf{p}}$. We thus take $Z = \mathrm{dist}(\hat{f})$ and substitute Eq. (B.2) into Eq. (B.4), giving us a simplified version of Eq. (B.3):

$$E_{\alpha f}(x, y) = \frac{\left(\frac{F}{\alpha}\right)^2 \cos\theta}{z^2\, b_{\alpha f}^2} \int_{\mathbf{q} \in \mathcal{C}_{xy}(\alpha,f)} L(\mathbf{q}(\omega), \omega) \cos^2\beta(\omega) \cos\gamma(\omega)\, \mathrm{d}\mathbf{q} \ , \tag{B.5}$$

where the blur diameter $b_{\alpha f}$ is what we hold fixed, and the only remaining terms that depend on lens setting are $\beta(\omega)$ and $\gamma(\omega)$. But from condition **C2**, both $\beta(\omega)$ and $\gamma(\omega)$ will be constant over all $(\alpha, f)$. Therefore, the contribution of a differential scene patch $\mathrm{d}\mathbf{q}$ to image irradiance is constant over all lens settings corresponding to the same blur diameter.

The only remaining issue concerns the domain of integration for Eq. (B.5), *i.e.*, the scene surface intersected by $\mathcal{C}_{xy}(\alpha, f)$, which varies in general with lens setting. However, given approximately constant depth at the boundary of the blur circle, as implied by condition **C5**, this

domain will be constant as well.

In practice, equi-blur constancy can actually tolerate significant depth variation within the blur circle, because such variations will be averaged over the defocused region of the scene.

# Appendix C

# Analytic Gradients for Layer-Based Restoration

Because our image formation model is a composition of linear operators plus clipping, the gradients of the objective function defined in Eqs. (4.8)–(4.9) take a compact analytic form.

Intuitively, our image formation model can be thought of as spatially-varying linear filtering, analogous to convolution ("distributing" image intensity according to the blur diameters and layering). Thus, the adjoint operator that defines its gradients corresponds to spatially-varying linear filtering as well, analogous to correlation ("gathering" image intensity) [106].

**Simplified gradient formulas.** For clarity, we first present gradients of the objective function assuming a single aperture, $a$, without inpainting:

$$\frac{\partial \mathcal{O}}{\partial \mathbf{L}} = e_a \mathbf{U}_a \sum_{k=1}^{K} \left[ \mathbf{A}_k \mathbf{M}_k \Delta \star B_{\sigma_k} \right] + \frac{\partial \|\mathbf{L}\|_\beta}{\partial \mathbf{L}} \tag{C.1}$$

$$\frac{\partial \mathcal{O}}{\partial \sigma_k} = e_a \mathbf{U}_a \sum_{x,y} \left[ \sum_{j=1}^{K} \left[ \mathbf{A}_j \mathbf{M}_j \Delta \star \frac{\partial B_{\sigma_j}}{\partial \sigma_j} \right] \right] \mathbf{A}_k \mathbf{L} \ , \tag{C.2}$$

where $\star$ denotes 2D correlation, and the binary mask

$$\mathbf{U}_a = \left[ e_a \overline{\mathbf{L}} < 1 \right] \tag{C.3}$$

indicates which pixels in the synthesized input image are unsaturated, thereby assigning zero gradients to over-saturated pixels. This definition resolves the special case where $e_a \overline{\mathbf{L}} = 1$ exactly, at which point the gradient of Eq. (4.9) is discontinuous. Since all matrix multiplications above are pixel-wise, we have omitted the operator $\cdot$ for brevity.

The only expression left to specify is the gradient for the regularization term in Eq. (4.11):

$$\frac{\partial \|\mathbf{L}\|_\beta}{\partial \mathbf{L}} \;=\; -\operatorname{div}\left(\frac{w(\mathbf{L})^2\,\nabla\mathbf{L}}{\sqrt{\big(w(\mathbf{L})\,\|\nabla\mathbf{L}\|\big)^2+\beta}}\right)\;, \tag{C.4}$$

where div is the divergence operator. This formula is a slight generalization of a previous treatment for the total variation norm [116], but it incorporates per-pixel weights, $w(\mathbf{L})$, to account for high dynamic range.

**Multiple aperture settings.** The generalization to the multiple aperture settings is straightforward. We add an outer summation over aperture, and relate blur diameter across aperture using scale factors that follow from Eq. (4.3), $s_a = \frac{D_a}{D_A}$. See footnote 3 (p. 85) for more detail about how we compute these scale factors in practice.

**Inpainting.** To generalize the gradient formulas to include inpainting, we assume that the inpainting operator for each layer $k$,

$$\mathcal{I}_k[\,\mathbf{L}\,] = \mathbf{A}'_k\mathbf{L} + \mathbf{A}''_k\mathbf{L}''_k\;, \tag{C.5}$$

can be expressed as a linear function of radiance. This model covers many existing inpainting methods, including choosing the nearest unoccluded pixel, PDE-based diffusion [21], and exemplar-based inpainting.

To compute the gradient, we need to determine the adjoint of the inpainting operator, $\mathcal{I}_k^\dagger[\cdot]$, which has the effect of "gathering" the inpainted radiance from its occluded destination and "returning" it to its unoccluded source. In matrix terms, if the inpainting operator is written as a large matrix left-multiplying the flattened scene radiance, $\boldsymbol{\mathcal{I}}_k$, the adjoint operator is simply its transpose, $\boldsymbol{\mathcal{I}}_k^T$.

**Gradient formulas.** Putting everything together, we obtain the final gradients:

$$\frac{\partial \mathcal{O}}{\partial \mathbf{L}} \;=\; \sum_{a=1}^{A} e_a \mathbf{U}_a\left[\sum_{k=1}^{K}\mathcal{I}_k^\dagger\left[\mathbf{A}'_k\mathbf{M}_k\Delta_a \;\star\; B_{(s_a\sigma_k)}\right]\right] + \frac{\partial\|\mathbf{L}\|_\beta}{\partial\mathbf{L}} \tag{C.6}$$

$$\frac{\partial \mathcal{O}}{\partial \sigma_k} \;=\; \sum_{a=1}^{A} s_a e_a \mathbf{U}_a\left[\sum_{x,y}\left[\sum_{j=1}^{K}\mathcal{I}_k^\dagger\left[\mathbf{A}'_j\mathbf{M}_j\Delta_a \;\star\; \frac{\partial B_{(s_a\sigma_j)}}{\partial(s_a\sigma_j)}\right]\right]\mathbf{A}'_k\mathbf{L}\right]\;. \tag{C.7}$$

# Appendix D

# Light-Efficiency Proofs

Theorem 1 follows as a consequence of Lemma 1 and four additional lemmas, while proving Theorem 2 is more direct. We first state Lemmas 2–5 and prove them below before addressing the theorems.

**Lemma 2** (Efficiency of Sequences with Sequential DOFs). *For every sequence $\mathcal{S}$, there is a sequence $\mathcal{S}'$ with sequential DOFs that spans the same synthetic DOF and has a total exposure time no larger than that of $\mathcal{S}$.*

**Lemma 3** (Permutation of Sequences with Sequential DOFs). *Given the left endpoint, $\alpha$, every permutation of $D_1, \ldots, D_n$ defines a capture sequence with sequential DOFs that has the same synthetic depth of field and the same total exposure time.*

**Lemma 4** (Optimality of Maximizing the Number of Photos). *Among all sequences with up to $n$ tuples whose synthetic DOF is $[\alpha, \beta]$, the sequence that minimizes total exposure time has exactly $n$ of them.*

**Lemma 5** (Optimality of Equal-Aperture Sequences). *If $\beta < (7+4\sqrt{3})\alpha$, then among all capture sequences with $n$ tuples whose synthetic DOF is $[\alpha, \beta]$, the sequence that minimizes total exposure time uses the same aperture for all tuples. Furthermore, this aperture is equal to*

$$D(n) = c\,\frac{\sqrt[n]{\beta} + \sqrt[n]{\alpha}}{\sqrt[n]{\beta} - \sqrt[n]{\alpha}}\ . \tag{D.1}$$

**Proof of Lemma 1.** We proceed inductively, by defining photo tuples whose DOFs "tile" the interval $[\alpha, \beta]$ from left to right. For the base case, the left endpoint of the first tuple's DOF

must be $\alpha_1 = \alpha$. Now consider the $i$-th tuple. Eq. (5.5) implies that the left endpoint $\alpha_i$ and the aperture diameter $D_i$ determine the DOF's right endpoint uniquely:

$$\beta_i = \frac{D_i + c}{D_i - c}\,\alpha_i \ . \tag{D.2}$$

The tuple's focus setting in Eq. (5.11) now follows by applying Eq. (5.6) to the interval $[\alpha_i, \beta_i]$. Finally, since the DOFs of tuple $i$ and $i+1$ are sequential, we have $\alpha_{i+1} = \beta_i$. □

**Proof of Lemma 2.** Let $\langle\ D,\ \tau,\ v\ \rangle$ be a tuple in $\mathcal{S}$, and let $[\alpha_1, \beta_1]$ be its depth of field. Now suppose that $\mathcal{S}$ contains another tuple whose depth of field, $[\alpha_2, \beta_2]$, overlaps with $[\alpha_1, \beta_1]$. Without loss of generality, assume that $\alpha_1 < \alpha_2 < \beta_1 < \beta_2$. We now replace $\langle\ D,\ \tau,\ v\ \rangle$ with a new tuple $\langle\ D',\ \tau',\ v'\ \rangle$ whose DOF is $[\alpha_1, \alpha_2]$ by setting $D'$ according to Eq. (5.5) and $v'$ according to Eq. (5.6). Since the DOF of the new tuple is narrower than the original, we have $D' > D$ and, hence, $\tau' < \tau$. Note that this tuple replacement preserves the synthetic DOF of the original sequence. We can apply this construction repeatedly until no tuples exist with overlapping DOFs. □

**Proof of Lemma 3.** From Eq. (5.11) it follows that the total exposure time is

$$\tau = \sum_{i=1}^{n} \frac{L^*}{{D_i}^2}\ , \tag{D.3}$$

which is invariant to the permutation. To show that the synthetic DOF is also permutation invariant, we apply Eq. (D.2) recursively $n$ times to obtain the right endpoint of the synthetic DOF:

$$\beta_n = \alpha \prod_{i=1}^{n} \frac{D_i + c}{D_i - c}\ . \tag{D.4}$$

It follows that $\beta_n$ is invariant to the permutation. □

**Proof of Lemma 4.** From Lemma 2 it follows that among all sequences up to length $n$ whose DOF is $[\alpha, \beta]$, there is a sequence $\mathcal{S}^*$ with minimum total exposure time whose tuples have sequential DOFs. Furthermore, Lemmas 1 and 3 imply that this capture sequence is fully determined by a sequence of $n'$ aperture settings, $D_1 \le D_2 \le \cdots \le D_{n'}$, for some $n' \le n$. These settings partition the interval $[\alpha, \beta]$ into $n'$ sub-intervals, whose endpoints are given by Eq. (5.12):

$$\alpha = \alpha_1 < \overbrace{\alpha_2 < \cdots < \alpha_{n'}}^{\text{determined by } \mathcal{S}^*} < \beta_{n'} = \beta\ . \tag{D.5}$$

It therefore suffices to show that placing $n' - 1$ points in $[\alpha, \beta]$ is most efficient when $n' = n$. To do this, we show that splitting a sub-interval always produces a more efficient capture sequence.

Consider the case $n = 2$, where the sub-interval to be split is actually equal to $[\alpha, \beta]$. Let $x \in [\alpha, \beta]$ be a splitting point. The exposure time for the sub-intervals $[\alpha, x]$ and $[x, \beta]$ can be obtained by combining Eqs. (5.5) and (5.1):

$$\tau(x) = \frac{L}{c^2}\left(\frac{x-\alpha}{x+\alpha}\right)^2 + \frac{L}{c^2}\left(\frac{\beta-x}{\beta+x}\right)^2 , \tag{D.6}$$

Differentiating Eq. (D.6) and evaluating it for $x = \alpha$ we obtain

$$\left.\frac{d\tau}{dx}\right|_{x=\alpha} = -\frac{4L}{c^2}\frac{(\beta-\alpha)\,\beta}{(\beta+\alpha)^3} < 0 . \tag{D.7}$$

Similarly, it is possible to show that $\frac{d\tau}{dx}$ is positive for $x = \beta$. Since $\tau(x)$ is continuous in $[\alpha, \beta]$, it follows that the minimum of $\tau(x)$ occurs strictly inside the interval. Hence, splitting the interval always reduces total exposure time. The general case for $n$ intervals follows by induction. □

**Proof of Lemma 5.** As in the proof of Lemma 4, we consider the case where $n = 2$. From that lemma it follows that the most efficient sequence involves splitting $[\alpha, \beta]$ into two sub-intervals $[\alpha, x]$ and $[x, \beta]$. To prove Lemma 5 we now show that the optimal split corresponds to a sequence with two identical aperture settings. Solving for $\frac{d\tau}{dx} = 0$ we obtain four solutions:

$$x = \left\{ \pm\sqrt{\alpha\beta}\,,\ \frac{(8\alpha\beta + \Delta) \pm (\beta-\alpha)\sqrt{\Delta}}{2(\beta+\alpha)} \right\} , \tag{D.8}$$

where $\Delta = \alpha^2 - 14\alpha\beta + \beta^2$. The inequality condition of Lemma 5 implies that $\Delta < 0$. Hence, the only real and positive solution is $x = \sqrt{\alpha\beta}$. From Eq. (5.5) it now follows that the intervals $[\alpha, \sqrt{\alpha\beta}]$ and $[\sqrt{\alpha\beta}, \beta]$ both correspond to an aperture equal to $c\,\frac{\sqrt{\beta}+\sqrt{\alpha}}{\sqrt{\beta}-\sqrt{\alpha}}$. To prove the Lemma for $n > 2$, we replace the sum in Eq. (D.6) with a sum of $n$ terms corresponding to the sub-divisions of $[\alpha, \beta]$, and then apply the above proof to each endpoint of that subdivision. This generates a set of relations, $\{\alpha_i = \sqrt{\alpha_{i-1}\alpha_{i+1}}\}_{i=2}^{n}$, which combine to define Eq. (D.1) uniquely. □

**Proof Sketch of Theorem 1.** We proceed in four steps. First, we consider sequences whose synthetic DOF is equal to $[\alpha, \beta]$. From Lemmas 4 and 5 it follows that the most efficient sequence, $\mathcal{S}'$, among this set has diameter and length given by Eq. (5.9). Second, we show that sequences with a larger synthetic DOF that are potentially more efficient can have at most one

more tuple. Third, we show that the most efficient of these sequences, $\mathcal{S}''$, uses a single diameter equal to $D_{\max}$. Finally, the decision rule in Eq. (5.10) follows by comparing the total exposure times of $\mathcal{S}'$ and $\mathcal{S}''$.

**Proof of Theorem 1.** We first consider the most efficient capture sequence, $\mathcal{S}'$, among all sequences whose synthetic DOF is identical to $[\alpha, \beta]$. Lemmas 4 and 5 imply that the most efficient sequence (1) has maximal length and (2) uses the same aperture for all tuples. More specifically, consider such a sequence of $n$ photos with diameter $D_i = D(n)$, for all $i$, according to Eq. (D.1). This sequence satisfies Eq. (D.4) with $\beta_n = \beta$, and we can manipulate this equation to obtain:

$$n = \frac{\log \frac{\alpha}{\beta}}{\log\left(\frac{D(n)-c}{D(n)+c}\right)} \quad . \tag{D.9}$$

Note that while $n$ increases monotonically with aperture diameter, the maximum aperture diameter $D_{max}$ restricts the maximal $n$ for which such an even subdivision is possible. This maximal $n$, whose formula is provided by Eq. (5.9), can be found by evaluating Eq. (D.9) with an aperture diameter of $D_{max}$.

While $\mathcal{S}'$ is the most efficient sequence among those whose synthetic DOFs equal to $[\alpha, \beta]$, there may be sequences whose DOF strictly contains this interval that are even more efficient. We now seek the most efficient sequence, $\mathcal{S}''$ among this class. To find it, we use two observations. First, $\mathcal{S}''$ must have length at most $n + 1$. This is because longer sequences must include a tuple whose DOF lies entirely outside $[\alpha, \beta]$. Second, among all sequences of length $n + 1$, the most efficient sequence is the one whose aperture diameters are all equal to the maximum possible value, $D_{max}$. This follows from the fact that any choice of $n + 1$ apertures is sufficient to span the DOF, so the most efficient such choice involves the largest apertures possible.

From the above considerations it follows that the optimal capture sequence will be an equal-aperture sequence whose aperture will be either $D(n)$ or $D_{max}$. The test in Eq. (5.10) comes from comparing the total exposure times of the sequences $\mathcal{S}'$ and $\mathcal{S}''$ using Eq. (D.3). The theorem's inequality condition comes from Lemma 5. □

**Proof of Theorem 2.** The formulation of the integer linear program in Eqs. (5.13)–(5.16) follows in a straightforward fashion from our objective of minimizing total exposure time, plus the constraint that the apertures used in the optimal capture sequence must span the desired DOF.

First, note that the multiplicities $n_i$ are non-negative integers, since they correspond to the

number of photos taken with each discrete aperture $D_i$. This is expressed in Eqs. (5.15)–(5.16). Second, we can rewrite the total exposure time given by Eq. (D.3) in terms of the multiplicities:

$$\tau = \sum_{i=1}^{m} n_i \frac{L^*}{{D_i}^2} , \tag{D.10}$$

This corresponds directly to Eq. (5.13), and is linear in the multiplicities being optimized. Finally, we can rewrite the expression for the right endpoint of the synthetic DOF provided by Eq. (D.4) in terms of the multiplicities as well:

$$\beta_m = \alpha \prod_{i=1}^{m} \left( \frac{D_i + c}{D_i - c} \right)^{n_i} . \tag{D.11}$$

Because all sequences we consider are sequential, the DOF $[\alpha, \beta]$ will be spanned without any gaps provided that the right endpoint satisfies $\beta_m \geq \beta$. By combining this constraint with Eq. (D.11) and taking logarithms, we obtain the inequality in Eq. (5.14), which is linear in the multiplicities being optimized as well. □

# Bibliography


[1] OPTIS OpticsWorks, http://www.optis-world.com/.

[2] Canon lens specifications, http://www.usa.canon.com/eflenses/pdf/spec.pdf.

[3] Canon 1D Mark II noise analysis, http://www.clarkvision.com/imagedetail/evaluation-1d2/.

[4] Technical Innovations, http://www.robofocus.com/.

[5] Flickr HDR group, http://www.flickr.com/groups/hdr/.

[6] CHDK, http://chdk.wikia.com/.

[7] Canon MTF charts, http://software.canon-europe.com/files/documents/EF_Lens_Work_Book_10_EN.pdf.

[8] S. Abrahamsson, S. Usawa, and M. Gustafsson. A new approach to extended focus for high-speed, high-resolution biological microscopy. In *Proc. SPIE*, volume 6090, page 60900N, 2006.

[9] E. H. Adelson and J. Y. A. Wang. Single lens stereo with a plenoptic camera. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):99–106, 1992.

[10] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *Proc. ACM SIGGRAPH*, pages 294–302, 2004.

[11] M. Aggarwal and N. Ahuja. A pupil-centric model of image formation. *International Journal of Computer Vision*, 48(3):195–214, 2002.

[12] M. Aggarwal, H. Hua, and N. Ahuja. On cosine-fourth and vignetting effects in real lenses. In *Proc. International Conference on Computer Vision*, volume 1, pages 472–479, 2001.

[13] N. Ahuja and A. L. Abbott. Active stereo: Integrating disparity, vergence, focus, aperture and calibration for surface estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(10):1007–1029, Oct. 1993.

[14] K. Aizawa, K. Kodama, and A. Kubota. Producing object-based special effects by fusing multiple differently focused images. *IEEE Trans. on Circuits and Systems for Video Technology*, 10(2):323–330, Mar. 2000.

[15] N. Asada, H. Fujiwara, and T. Matsuyama. Edge and depth from focus. *International Journal of Computer Vision*, 26(2):153–163, 1998.

[16] N. Asada, H. Fujiwara, and T. Matsuyama. Seeing behind the scene: Analysis of photometric properties of occluding edges by the reversed projection blurring model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(2):155–167, 1998.

[17] M. Baba, N. Asada, A. Oda, and T. Migita. A thin lens based camera model for depth estimation from blur and translation by zooming. In *Proc. Vision Interface*, pages 274–281, 2002.

[18] S. Bae and F. Durand. Defocus magnification. In *Proc. Eurographics*, 2007.

[19] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–25, 2004.

[20] M. Ben Ezra and S. Nayar. Motion-based motion deblurring. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(6):689–698, June 2004.

[21] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. ACM SIGGRAPH*, pages 417–424, 2000.

[22] S. S. Bhasin and S. Chaudhuri. Depth from defocus in presence of partial self occlusion. In *Proc. International Conference on Computer Vision*, volume 2, pages 488–493, 2001.

[23] J.-Y. Bouguet. Camera calibration toolbox for Matlab (Oct. 14, 2004), `http://vision.caltech.edu/bouguetj/calib_doc/`, 2004.

[24] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, Sept. 2004.

[25] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov. 2001.

[26] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, Aug. 2007.

[27] J. Buzzi and F. Guichard. Uniqueness of blur measure. In *Proc. International Conference on Image Processing*, volume 5, pages 2985–2988, 2004.

[28] W. T. Cathey and E. R. Dowski. New paradigm for imaging systems. *Applied Optics*, 41(29):6080–6092, Oct. 2002.

[29] S. Chaudhuri. Defocus morphing in real aperture images. *J. Optical Society of America A*, 22(11):2357–2365, Nov. 2005.

[30] T. Darrell and K. Wohn. Pyramid based depth from focus. In *Proc. Computer Vision and Pattern Recognition*, pages 504–509, 1988.

[31] P. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. ACM SIGGRAPH*, pages 369–378, 1997.

[32] E. R. Dowski, Jr. and W. T. Cathey. Single-lens single-image incoherent passive-ranging systems. *Applied Optics*, 33(29):6762–6773, Oct. 1994.

[33] E. R. Dowski, Jr. and W. T. Cathey. Extended depth of field through wave-front coding. *Applied Optics*, 34(11):1859–1866, Apr. 1995.

[34] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.*, 23(3):673–678, 2004.

[35] J. Ens and P. Lawrence. An investigation of methods for determining depth from focus. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(2):97–108, 1993.

[36] H. Farid and E. P. Simoncelli. Range estimation by optical differentiation. *J. Optical Society of America A*, 15(7):1777–1786, 1998.

[37] P. Favaro. Shape from focus and defocus: Convexity, quasiconvexity and defocus-invariant textures. In *Proc. International Conference on Computer Vision*, pages 1–7, 2007.

[38] P. Favaro, A. Mennucci, and S. Soatto. Observing shape from defocused images. *International Journal of Computer Vision*, 52(1):25–43, 2003.

[39] P. Favaro, S. J. Osher, S. Soatto, and L. A. Vese. 3D shape from anisotropic diffusion. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 179–186, 2003.

[40] P. Favaro and S. Soatto. Shape and radiance estimation from the information divergence of blurred images. In *Proc. European Conference on Computer Vision*, volume 1, pages 755–768, 2000.

[41] P. Favaro and S. Soatto. Learning shape from defocus. In *Proc. European Conference on Computer Vision*, volume 2, pages 735–745, 2002.

[42] P. Favaro and S. Soatto. Seeing beyond occlusions (and other marvels of a finite lens aperture). In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 579–586, 2003.

[43] P. Favaro and S. Soatto. A geometric approach to shape from defocus. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):406–417, Mar. 2005.

[44] R. Fergus, A. Torralba, and W. T. Freeman. Random lens imaging. Technical Report MIT-CSAIL-TR-2006-058, MIT, 2006.

[45] A. Fitzgibbon, Y. Wexler, and A. Zisserman. Image-based rendering using image-based priors. *International Journal of Computer Vision*, 63(2):141–151, July 2005.

[46] C. S. Fraser and M. R. Shortis. Variation of distortion within the photographic field. *Photogrammetric Engineering and Remote Sensing*, 58(6):851–855, 1992.

[47] T. Georgiev, C. Zheng, S. Nayar, D. Salesin, B. Curless, and C. Intwala. Spatio-angular resolution tradeoffs in integral photography. In *Proc. Eurographics Symposium on Rendering*, pages 263–272, 2006.

[48] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Proc. ACM SIGGRAPH*, pages 43–54, 1996.

[49] P. Green, W. Sun, W. Matusik, and F. Durand. Multi-aperture photography. In *Proc. ACM SIGGRAPH*, 2007.

[50] M. D. Grossberg and S. K. Nayar. Modeling the space of camera response functions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(10):1272–1282, 2004.

[51] P. Haeberli. A multifocus method for controlling depth of field. Technical report, GRAFICA Obscura, Oct. 1994. `http://www.graficaobscura.com/depth/`.

[52] S. W. Hasinoff, S. B. Kang, and R. Szeliski. Boundary matting for view synthesis. *Computer Vision and Image Understanding*, 103(1):22–32, July 2006.

[53] S. W. Hasinoff and K. N. Kutulakos. Confocal stereo. In *Proc. European Conference on Computer Vision*, volume 1, pages 620–634, 2006.

[54] S. W. Hasinoff and K. N. Kutulakos. A layer-based restoration framework for variable-aperture photography. In *Proc. International Conference on Computer Vision*, pages 1–8, 2007.

[55] G. Häusler. A method to increase the depth of focus by two step image processing. *Optics Communications*, 6(1):38–42, 1972.

[56] G. E. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(3):267–276, 1994.

[57] A. Hertzmann and S. M. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1254–1264, 2005.

[58] S. Hiura and T. Matsuyama. Depth measurement by the multi-focus camera. In *Proc. Computer Vision and Pattern Recognition*, pages 953–959, 1998.

[59] H. H. Hopkins. The frequency response of a defocused optical system. *Proc. of the Royal Society of London, Series A*, 231(1184):91–103, 1955.

[60] B. K. P. Horn. Focusing. Technical Report AIM-160, Massachusetts Institute of Technology, 1968.

[61] A. Isaksen, L. McMillan, and S. J. Gortler. Dynamically reparameterized light fields. In *Proc. ACM SIGGRAPH*, pages 297–306, 2000.

[62] H. Jin and P. Favaro. A variational approach to shape from defocus. In *Proc. European Conference on Computer Vision*, volume 2, pages 18–30, 2002.

[63] S. B. Kang and R. S. Weiss. Can we calibrate a camera using an image of a flat, textureless Lambertian surface? In *Proc. European Conference on Computer Vision*, volume 2, pages 640–653, 2000.

[64] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1(3):223–237, 1987.

[65] A. Kubota and K. Aizawa. Inverse filters for reconstruction of arbitrarily focused images-from two differently focused images. In *Proc. International Conference on Image Processing*, volume 1, pages 101–104, 2000.

[66] A. Kubota, K. Kodama, and K. Aizawa. Registration and blur estimation methods for multiple differently focused images. In *Proc. International Conference on Image Processing*, volume 2, pages 447–451, 1999.

[67] A. Kubota, K. Takahashi, K. Aizawa, and T. Chen. All-focused light field rendering. In *Proc. Eurographics Symposium on Rendering*, 2004.

[68] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):197–216, July 2000.

[69] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *Proc. ACM SIGGRAPH*, 2007.

[70] A. Levin, W. T. Freeman, and F. Durand. Understanding camera trade-offs through a Bayesian analysis of light field projections. In *Proc. European Conference on Computer Vision*, 2008. to appear.

[71] A. Levin, W. T. Freeman, and F. Durand. Understanding camera trade-offs through a Bayesian analysis of light field projections – A revision. Technical report, CSAIL, MIT, July 2008.

[72] M. Levoy, May 2008. Personal communication.

[73] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. T. Bolas. Synthetic aperture confocal imaging. In *Proc. ACM SIGGRAPH*, pages 825–834, 2004.

[74] M. Levoy and P. Hanrahan. Light field rendering. In *Proc. ACM SIGGRAPH*, pages 31–42, 1996.

[75] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz. Light field microscopy. In *Proc. ACM SIGGRAPH*, pages 924–934, 2006.

[76] C. Liu, R. Szeliski, S. Kang, C. Zitnick, and W. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(2):299–314, Feb. 2008.

[77] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand. Defocus video matting. In *Proc. ACM SIGGRAPH*, pages 567–576, 2005.

[78] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *Proc. Computer Vision and Pattern Recognition*, pages 1374–1380, 1999.

[79] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. Active refocusing of images and videos. In *Proc. ACM SIGGRAPH*, 2007.

[80] H. Nair and C. Stewart. Robust focus ranging. In *Proc. Computer Vision and Pattern Recognition*, pages 309–314, 1992.

[81] S. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(12):1186–1198, Dec. 1996.

[82] S. K. Nayar and Y. Nakagawa. Shape from focus: an effective approach for rough surfaces. In *Proc. International Conference on Robotics and Automation*, volume 2, pages 218–225, 1990.

[83] R. Ng. Fourier slice photography. In *Proc. ACM SIGGRAPH*, pages 735–744, 2005.

[84] R. Ng and P. Hanrahan. Digital correction of lens aberrations in light field photography. In *Proc. SPIE*, volume 6342, 63421E, 2007.

[85] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CTSR 2005-02, Dept. Computer Science, Stanford University, 2005.

[86] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[87] J. Ogden, E. Adelson, J. R. Bergen, and P. Burt. Pyramid-based computer graphics. *RCA Engineer*, 30(5):4–15, 1985.

[88] J. Ojeda-Castaneda, E. Tepichin, and A. Pons. Apodization of annular apertures: Strehl ratio. *Applied Optics*, 27(24):5140–5145, Dec. 1988.

[89] S. Paris, H. Briceño, and F. Sillion. Capture of hair geometry from multiple images. In *Proc. ACM SIGGRAPH*, pages 712–719, 2004.

[90] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, May 2003.

[91] A. Pentland, T. Darrell, M. Turk, and W. Huang. A simple, real-time range camera. In *Proc. Computer Vision and Pattern Recognition*, pages 256–261, 1989.

[92] A. P. Pentland. A new sense for depth of field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(4):523–531, July 1987.

[93] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.*, 23(3):664–672, 2004.

[94] A. N. Rajagopalan and S. Chaudhuri. A variational approach to recovering depth from defocused images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(10):1158–1164, Oct. 1997.

[95] A. N. Rajagopalan and S. Chaudhuri. An MRF model-based approach to simultaneous recovery of depth and restoration from defocused images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(7):577–589, July 1999.

[96] R. Raskar, A. Agrawal, and J. Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *Proc. ACM SIGGRAPH*, pages 795–804, 2006.

[97] R. Redondo, F. Sroubek, S. Fischer, and G. Cristóbal. Multifocus fusion with multisize windows. In *Proc. SPIE*, volume 5909, pages 380–388, 2005.

[98] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, Nov. 1994.

[99] Y. Y. Schechner and N. Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162, Sept. 2000.

[100] Y. Y. Schechner, N. Kiryati, and R. Basri. Separation of transparent layers using focus. *International Journal of Computer Vision*, 39(1):25–39, 2000.

[101] Y. Y. Schechner and S. K. Nayar. Generalized mosaicing: High dynamic range in a wide field of view. *International Journal of Computer Vision*, 53(3):245–267, 2004.

[102] S.-W. Shih, P.-S. Kao, and W.-S. Guo. Error analysis and accuracy improvement of depth from defocusing. In *Proc. IPPR Conference on Computer Vision, Graphics and Image Processing*, volume 4, pages 250–257, 2003.

[103] M. Shimizu and M. Okutomi. Reflection stereo – novel monocular stereo using a transparent plate. In *Proc. Canadian Conference on Computer and Robot Vision*, pages 14–21, 2006.

[104] A. Smith and J. Blinn. Blue screen matting. In *Proc. ACM SIGGRAPH*, pages 259–268, 1996.

[105] W. J. Smith. *Modern Optical Engineering*. McGraw-Hill, New York, 3rd edition, 2000.

[106] M. Šorel. *Multichannel blind restoration of images with space-variant degradations*. PhD thesis, Charles University in Prague, Dept. of Software Engineering, 2007.

[107] M. Šorel and J. Flusser. Simultaneous recovery of scene structure and blind restoration of defocused images. In *Proc. Computer Vision Winter Workshop*, pages 40–45, 2006.

[108] F. Šroubek, S. Gabarda, R. Redondo, S. Fischer, and G. Cristóbal. Multifocus fusion with oriented windows. In *Proc. SPIE*, volume 5839, pages 264–273, 2005.

[109] M. Subbarao. Parallel depth recovery by changing camera parameters. In *Proc. International Conference on Computer Vision*, pages 149–155, 1988.

[110] M. Subbarao and N. Gurumoorthy. Depth recovery from blurred edges. In *Proc. Computer Vision and Pattern Recognition*, pages 498–503, 1988.

[111] M. Subbarao and G. Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, Dec. 1994.

[112] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA, USA, 2005.

[113] J. Telleen, A. Sullivan, J. Yee, P. Gunawardane, O. Wang, I. Collins, and J. Davis. Synthetic shutter speed imaging. In *Proc. Eurographics*, pages 591–598, 2007.

[114] V. Vaish, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 2331–2338, 2006.

[115] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *Proc. ACM SIGGRAPH*, 2007.

[116] C. Vogel and M. Oman. Fast, robust total variation based reconstruction of noisy, blurred images. *IEEE Trans. on Image Processing*, 7(6):813–824, June 1998.

[117] W. Wallace, L. H. Schaefer, and J. R. Swedlow. A workingperson's guide to deconvolution in light microscopy. *Biotechniques*, 31(5):1076–1097, 2001.

[118] M. Watanabe and S. K. Nayar. Minimal operator set for passive depth from defocus. In *Proc. Computer Vision and Pattern Recognition*, pages 431–438, 1996.

[119] M. Watanabe and S. K. Nayar. Telecentric optics for focus analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(12):1360–1365, Dec 1997.

[120] M. Watanabe and S. K. Nayar. Rational filters for passive depth from defocus. *International Journal of Computer Vision*, 27(3):203–225, 1998.

[121] R. H. Webb. Confocal optical microscopy. *Reports on Progress in Physics*, 59(3):427–471, 1996.

[122] Y. Wei, E. Ofek, L. Quan, and H.-Y. Shum. Modeling hair from multiple views. In *Proc. ACM SIGGRAPH*, pages 816–820, 2005.

[123] W. T. Welford. Use of annular apertures to increase focal depth. *J. Optical Society of America A*, 50(8):749–753, Aug. 1960.

[124] R. Willson. Modeling and calibration of automated zoom lenses. In *Proc. SPIE*, volume 2350: Videometrics III, pages 170–186, Oct 1994.

[125] R. Willson. *Modeling and Calibration of Automated Zoom Lenses*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, January 1994.

[126] R. Willson and S. Shafer. Dynamic lens compensation for active color imaging and constant magnification focusing. Technical Report CMU-RI-TR-91-26, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Nov 1991.

[127] R. Willson and S. Shafer. What is the center of the image? *J. Optical Society of America A*, 11(11):2946–2955, Nov 1994.

[128] Y. Xiong and S. Shafer. Depth from focusing and defocusing. In *Proc. Computer Vision and Pattern Recognition*, pages 68–73, June 1993.

[129] Y. Xiong and S. Shafer. Moment and hypergeometric filters for high precision computation of focus, stereo and optical flow. *International Journal of Computer Vision*, 22(1):25–59, Feb. 1997.

[130] N. Xu, K. Tan, H. Arora, and N. Ahuja. Generating omnifocus images using graph cuts and a new focus measure. In *Proc. International Conference on Pattern Recognition*, volume 4, pages 697–700, 2004.

[131] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum. Image deblurring with blurred/noisy image pairs. In *Proc. ACM SIGGRAPH*, 2007.

[132] L. Zhang and S. K. Nayar. Projection defocus analysis for scene capture and image display. In *Proc. ACM SIGGRAPH*, pages 907–915, 2006.

[133] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *Proc. ACM SIGGRAPH*, pages 548–558, Aug. 2004.

[134] N. F. Zhang, M. T. Postek, R. D. Larrabee, A. E. Vladár, W. J. Keery, and S. N. Jones. Image sharpness measurement in the scanning electron microscope—Part III. *Scanning*, 21:246–252, 1999.

[135] Z. Zhang and R. Blum. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proc. of the IEEE*, 87(8):1315–1326, Aug. 1999.

[136] T. E. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. *International Journal of Computer Vision*, 49(2-3):215–227, 2002.

[137] C. L. Zitnick and S. B. Kang. Stereo for image-based rendering using image oversegmentation. *International Journal of Computer Vision*, 75(1):49–65, 2007.

[138] A. Zomet and S. K. Nayar. Lensless imaging with a controllable aperture. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 339–346, June 2006.