Int J Comput Vis (2009) 81: 82–104
DOI 10.1007/s11263-008-0164-2

# Confocal Stereo

**Samuel W. Hasinoff · Kiriakos N. Kutulakos**

Received: 25 October 2006 / Accepted: 29 July 2008 / Published online: 10 September 2008


**Abstract** 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 (12MP or more) with large-aperture lenses (e.g., f1.2). 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.

**Keywords** Defocus · Depth from focus · Depth from defocus · 3D reconstruction · Stereo · Camera calibration · Wide-aperture lenses

## 1 Introduction

Recent years have seen many advances in the problem of reconstructing complex 3D scenes from multiple photographs (Zitnick et al. 2004; Fitzgibbon et al. 2005; Hertzmann and Seitz 2005). Despite this progress, however, there are many common scenes for which obtaining detailed 3D models is beyond the state of the art. One such 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 (Fitzgibbon et al. 2005), require a strong prior model for the scene (Wei et al. 2005; Paris et al. 2004), 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 12 megapixels (MP), even individual strands of hair may be one or more pixels wide (Fig. 1a, b). In this paper, we explore the possibility of reconstructing such scenes with 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

Part of this work was done while the authors were visiting Microsoft Research Asia, in the roles of research intern and Visiting Scholar respectively.

S.W. Hasinoff (✉) · K.N. Kutulakos
Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada
e-mail: hasinoff@cs.toronto.edu

K.N. Kutulakos
e-mail: kyros@cs.toronto.edu

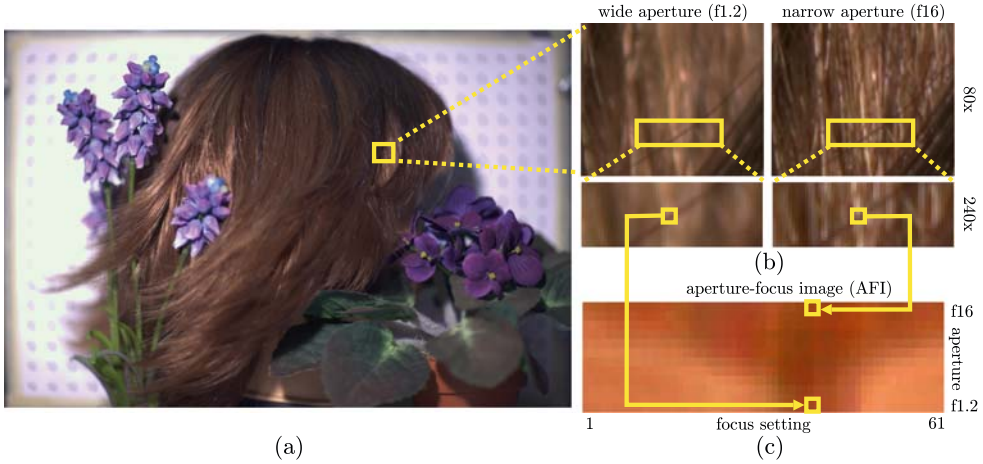


**Fig. 1** (**a**) Wide-aperture image of a complex scene. (**b**) *Left*: Successive close-ups of a region in (**a**), showing a single in-focus strand of hair. *Right*: Narrow-aperture image of the same region, with everything in focus. Confocal constancy tells us that the intensity of in-focus pixels (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

camera and a laptop. The only key requirement is the ability to actively control both the aperture and focus setting 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., f1.2). 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 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 article by discussing the relation of this work to current approaches for reconstructing scenes that exploit defocus in wide-aperture images. Section 3 describes our generic imaging model and introduces the property of confocal constancy. Section 4 gives a brief overview of how we exploit this property for reconstruction and Sects. 5–6 discuss the radiometric and geometric calibration required to relate high resolution images taken with different lens settings. In Sect. 7 we show how the AFI for each pixel can be analyzed independently to estimate depth, using both confocal constancy and its generalization. Finally, Sect. 8 presents experimental results using images of complex real scenes, and one scene for which ground truth has been recovered.

## 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 (Krotkov 1987; Darrell and Wohn 1988; Nair and Stewart 1992). 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 (Pentland 1987; Subbarao and Surya 1994; Xiong and Shafer 1997; Watanabe and Nayar 1998; Favaro and Soatto 2005; Green et al. 2007). 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 (Levin et al. 2007; Veeraraghavan et al. 2007). 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 (Pentland 1987; Subbarao and Surya 1994; Green et al. 2007). 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 (Farid and Simoncelli 1998), illuminating the connection between defocus-based methods and small-baseline stereo (Farid and Simoncelli 1998; Schechner and Kiryati 2000). 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 (Rajagopalan and Chaudhuri 1999; Jin and Favaro 2002; Favaro et al. 2003a, 2003b; Bhasin and Chaudhuri 2001; Favaro and Soatto 2003; McGuire et al. 2005; Hasinoff and Kutulakos 2007). Some of the recent methods in this framework model defocus in greater detail to better handle occlusion boundaries (Bhasin and Chaudhuri 2001; Favaro and Soatto 2003; McGuire et al. 2005; Hasinoff and Kutulakos 2007), 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 (Levin et al. 2007; Veeraraghavan et al. 2007), 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 (Nayar et al. 1996; Favaro et al. 2003a; Moreno-Noguer et al. 2007). 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 (Webb 1996). This approach has also been explored for larger-scale opaque scenes (Levoy et al. 2004).

Most recently, Zhang and Nayar developed an active illumination method that also computes depth maps at sensor resolution (Zhang and Nayar 2006). 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 (12MP 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 (Subbarao and Surya 1994; Pentland 1987), or by fitting a global parametric model such as a cosine-fourth falloff (Kang and Weiss 2000). 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 (Debevec and Malik 1997; Grossberg and Nayar 2004) as well as slight temporal variations in ambient lighting. Furthermore, most methods for modeling geometric lens distortions due to changing focus

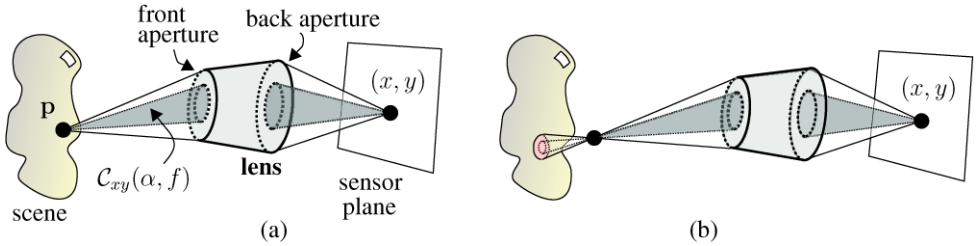


**Fig. 2** Generic lens model. (**a**) At the perfect focus setting of pixel $(x, y)$, the lens collects outgoing radiance from a scene point **p** and directs it toward the pixel. The 3D position of point **p** is uniquely determined by pixel $(x, y)$ and its perfect focus setting. The *shaded cone* 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

or zoom setting rely on simple magnification (Asada et al. 1998a; Darrell and Wohn 1988; Watanabe and Nayar 1997; Nayar et al. 1996) or radial distortion models (Willson 1994a), 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 (Levoy and Hanrahan 1996). Various such systems have been proposed in recent years, including camera arrays (Levoy and Hanrahan 1996; Isaksen et al. 2000), virtual camera arrays simulated using mirrors (Levoy et al. 2004), and arrays of lenslets in front of a standard imaging sensor (Adelson and Wang 1992; Ng 2005). 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 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, $I_{\alpha f}(x, y)$, where $I_{\alpha f}$ 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 (Krotkov 1987; Pentland 1987; Zhang and Nayar 2006).

Suppose that a 3D point **p** on an opaque surface is in perfect focus in image $I_{\alpha f}$ and suppose that it projects to pixel $(x, y)$. In this case, the light reaching the pixel is restricted to a cone from **p** that is determined by the aperture setting (Fig. 2). For a sensor with a linear response, the intensity $I_{\alpha f}(x, y)$ at the pixel is proportional to the integral of outgoing radiance over the cone, i.e.,

$$I_{\alpha f}(x, y) = \kappa \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} L(\mathbf{p}, \omega)\, \mathrm{d}\omega, \tag{1}$$

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 (Debevec and Malik 1997; Grossberg and Nayar 2004), 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, $\mathcal{C}_{xy}(\alpha_1, f) \subset \cdots \subset \mathcal{C}_{xy}(\alpha_A, f)$, 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 (1) disappears and the intensity for aperture $\alpha$ is proportional to the solid angle subtended by the associated cone, i.e.,

$$I_{\alpha f}(x, y) = \kappa \|\mathcal{C}_{xy}(\alpha, f)\| L(\mathbf{p}), \tag{2}$$

[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.

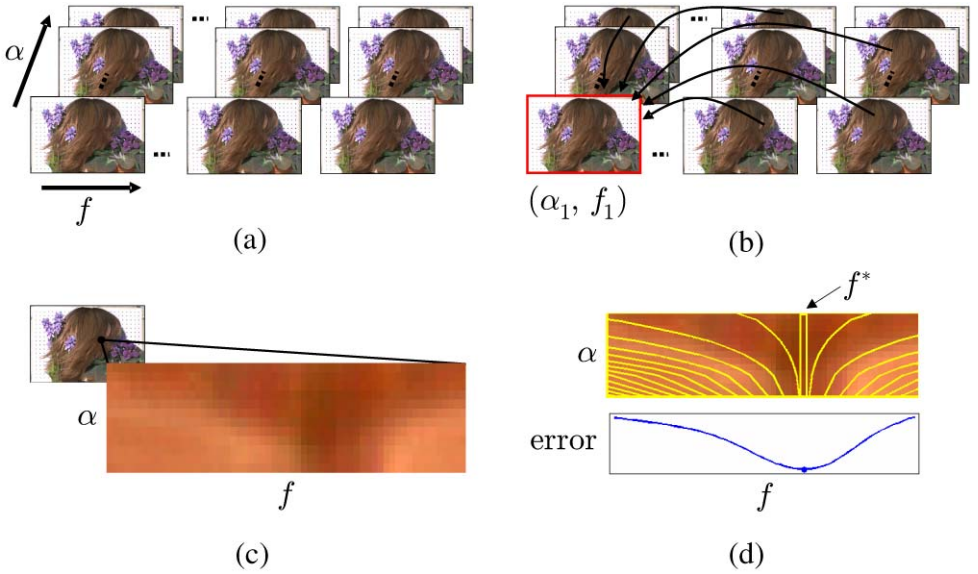


**Fig. 3** Overview of confocal stereo: (**a**) Acquire $A \times F$ images over $A$ apertures and $F$ focus settings. (**b**) Align all images to the reference image, taking into account both radiometric calibration (Sect. 5) and geometric distortion (Sect. 6). (**c**) Build the $A \times F$ aperture-focus image (AFI) for each pixel. (**d**) Process the AFI to find the best in-focus setting (Sect. 7)

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**

$$\frac{I_{\alpha f}(x, y)}{I_{\alpha_1 f}(x, y)} = \frac{\|\mathcal{C}_{xy}(\alpha, f)\|}{\|\mathcal{C}_{xy}(\alpha_1, f)\|} \stackrel{\text{def}}{=} R_{xy}(\alpha, f). \tag{3}$$

Intuitively, the constant of proportionality, $R_{xy}(\alpha, f)$, 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**.

[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.

## 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 $I_{\alpha f}(x, y)$ for different values of aperture $\alpha$ and focus $f$. This leads to the following reconstruction procedure (Fig. 3):

1. (*Relative exitance estimation*) Compute the relative exitance of the lens for the $A$ apertures and $F$ focus settings (Sect. 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 (Sect. 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 (Sect. 7).

## 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. 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

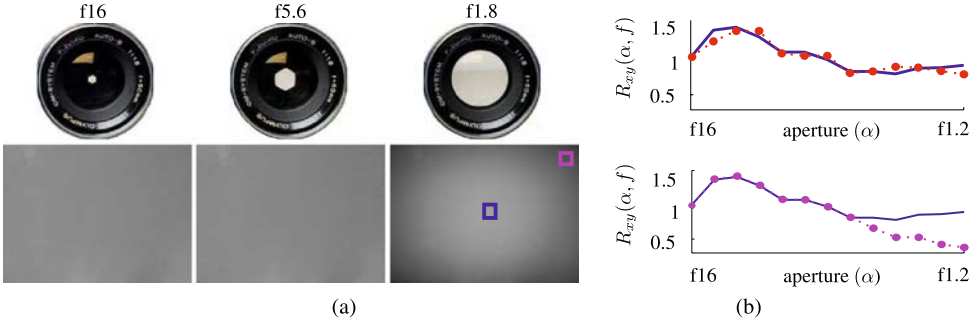


**Fig. 4** (**a**) Images of an SLR lens showing variation in aperture shape with corresponding images of a diffuse plane. (**b**) *Top*: comparison of relative exitances for the central pixel indicated in (**a**), as measured using (3) (*solid graph*), and as approximated using the f-stop values (*dotted*) according to $R_{xy}(\alpha, f) = \alpha_1^2/\alpha^2$ (Debevec and Malik 1997). *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

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 (Smith 2000; Kang and Weiss 2000) (Fig. 4b).

To deal with these issues, we explicitly compute the relative exitance of the lens, $R_{xy}(\alpha, f)$, 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 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 (3) to the luminance values of each pixel $(x, y)$ to recover $R_{xy}(\alpha_i, f)$. To obtain $R_{xy}(\alpha_i, f)$ 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 (3) assumes that pixel intensity is a linear function of radiance, we linearize the images using the inverse of the sensor response function, which we recover using standard techniques from the high dynamic range literature (Debevec and Malik 1997; Grossberg and Nayar 2004).

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, (1), and does not affect the focusing behavior of the lens.[4] Thus, we can fold variation in exposure time into the calculation of $R_{xy}(\alpha_i, f)$, 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 intervals 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.

## 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

[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%.

[4] A side-effect of manipulating the exposure time is that noise characteristics will change with varying intensity (Healey and Kondepudy 1994), however this phenomenon does not appear to be significant in our experiments.

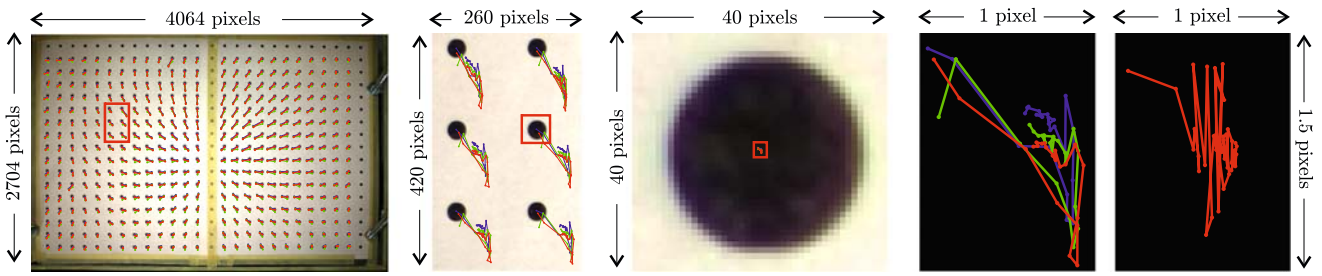


**Fig. 5** (Color online) (**a**–**e**) To evaluate stochastic lens distortions, we computed centroids of dot features for images of a static calibration pattern. (**a**–**d**) Successive close-ups of a centroid's trajectory for three cycles (*red*, *green*, *blue*) of the 23 aperture settings. In (**a**–**b**) 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

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. 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 12MP 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 (Darrell and Wohn 1988; Nayar et al. 1996; Favaro and Soatto 2002, 2003; Favaro et al. 2003b; Asada et al. 1998b) is insufficient to account for observed distortions.

### 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:

$$\mathbf{w}_f^{\mathrm{D}}(x,y) = \big[\, m_f + m_f(f-f_1)(k_0 + k_1 r + k_2 r^2) - 1 \,\big] \times \big[(x,y) - (x_c,y_c)\big], \tag{4}$$

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.

[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 $f$.

It is a standard procedure in many methods (Willson 1994a; Kubota et al. 2004) to model radial distortion using a polynomial of the radial displacement, $r$. A difference in our model is that the quadratic distortion term in (4) incorporates a linear dependence on the focus setting as well, consistent with more detailed calibration methods involving distortion components related to distance (Fraser and Shortis 1992). In our empirical tests, we have found that this term is necessary to obtain sub-pixel registration at high resolutions.

### 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. 5e). While these motions are clearly stochastic, we also observed a reproducible, aperture-dependent misalignment of about the same magnitude (Fig. 5a–d). In order to achieve sub-pixel alignment, we approximate these motions by a global 2D translation, estimated independently for every image:

$$\mathbf{w}_{\alpha f}^{\mathrm{S}}(x,y) = \mathbf{t}_{\alpha f}. \tag{5}$$

We observed these motions with two different Canon lenses and two 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 internal lens motions that are correlated with the action of the mechanical aperture.

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.

### 6.3 Offline Geometric Lens Calibration

We recover the complete distortion model of (4–5) 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:

$$E(x_c, y_c, \mathbf{m}, \mathbf{k}, \mathbf{T}) = \sum_{(x,y)} \sum_{f} \|\mathbf{w}^{\mathrm{D}}_{f}(x,y) + \mathbf{w}^{\mathrm{S}}_{\alpha_1 f}(x,y) - \Delta_{\alpha_1 f}(x,y)\|^2, \quad (6)$$

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 (1994) 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. 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 (Willson 1994b). To increase robustness to outliers, we run the optimization for (6) iteratively, removing features whose reprojection error is more than 3.0 times the median.

### 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 (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. 1a, for example).

We deal with this issue by using Lucas-Kanade registration to compute the residual stochastic translations in an image-based fashion (Darrell and Wohn 1988; Baker and Matthews 2004). 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.

## 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. 6a):

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

$$AFI_{xy}(\alpha, f) = \frac{1}{R_{xy}(\alpha, f)} \hat{I}_{\alpha f}(x,y), \quad (7)$$

where $\hat{I}_{\alpha f}$ denotes the images after global lighting correction (Sect. 5) and geometric image alignment (Sect. 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 $AFI(\alpha, f)$ to denote its AFI.

### 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 (7) already corrects for these variations, confocal constancy at a hypothesis $\hat{f}$ implies constant intensity within column $\hat{f}$ of

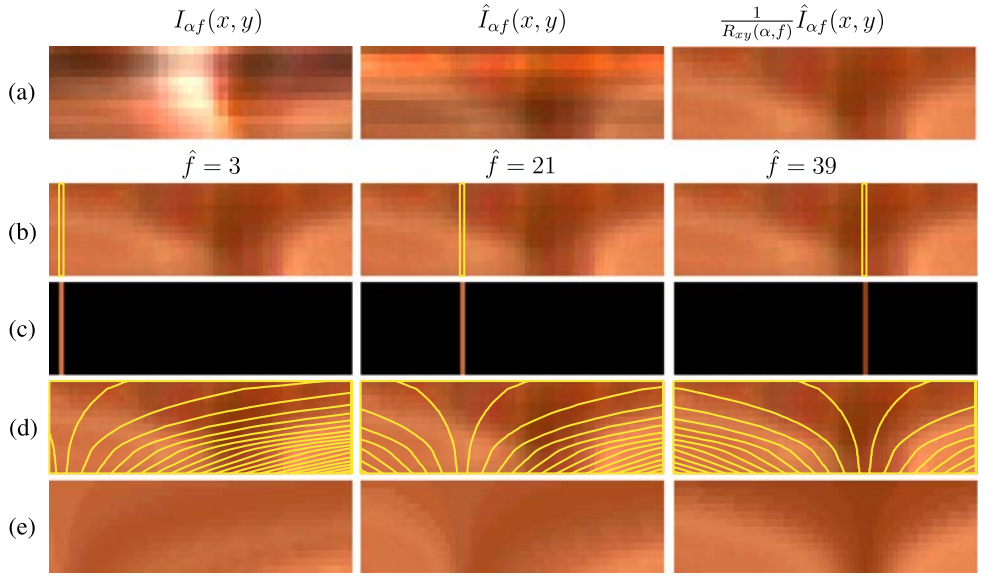


**Fig. 6** (**a**) The $A \times F$ measurements for the pixel shown in Fig. 1. *Left*: prior to image alignment. *Middle*: after image alignment. *Right*: after accounting for relative exitance (7). Note that the AFI's smooth structure is discernible only after both corrections. (**b**) Direct evaluation 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)

the AFI (Fig. 6b, c). Hence, to find the perfect focus setting we can simply find the column with minimum variance:

$$f^* = \arg\min_{\hat{f}} \mathrm{Var}\big\{AFI(1,\hat{f}),\ldots,AFI(A,\hat{f})\big\}. \tag{8}$$

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. 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.

### 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 (8) (Fig. 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 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. 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. 6e):

$$f^* = \arg\min_{\hat{f}} \sum_{i=1}^{F} \Big( w_i^{\hat{f}}\ \mathrm{Var}\big\{AFI(\alpha,f) \mid (\alpha,f) \in \mathcal{B}_i^{\hat{f}}\big\}\Big), \tag{9}$$

[6] While not analyzed in the context of confocal constancy or the AFI, this is a key observation exploited by depth from defocus approaches (Pentland 1987; Subbarao and Surya 1994; Farid and Simoncelli 1998; Watanabe and Nayar 1998; Favaro and Soatto 2003, 2005).

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 $w_i^{\hat{f}} = \text{area}(\mathcal{B}_i^{\hat{f}})$. For color images, as in (8), we compute the focus criterion for each RGB channel independently and then sum over channels.

To implement (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. 7a). Now consider "defocusing" the lens by changing its focus to $f$ (Fig. 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

$$b_{\alpha f} = \frac{\lambda}{\alpha} \frac{|\text{dist}(\hat{f}) - \text{dist}(f)|}{\text{dist}(f)}, \tag{10}$$

where $\lambda$ 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 (Smith 2000).

Given a focus hypothesis $\hat{f}$, (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. 6d and 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.

Equation (10) fully specifies our AFI model, and we have found that this model matches the observed pixel variations quite well in practice (Fig. 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 (Sects. 5–6), the equi-blur regions of the AFI are well-approximated by the equi-blur regions predicted by the thin-lens model (Smith 2000; Asada et al. 1998b). 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 Sect. 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.

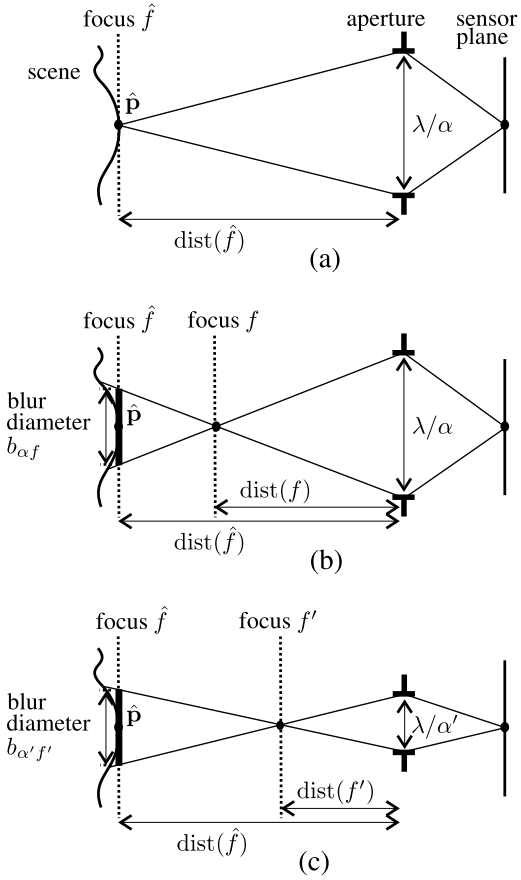


**Fig. 7** Quantifying the blur due to an out-of-focus setting. (**a**) At focus setting $\hat{f}$, scene point $\hat{\mathbf{p}}$ is in perfect focus. The aperture's effective diameter can be expressed in terms of its f-stop value $\alpha$ and the focal length $\lambda$. (**b**) For an out-of-focus setting $f$, we can use (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

## 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 16MP Canon EOS-1Ds Mark II ("box" dataset), and the 12MP 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

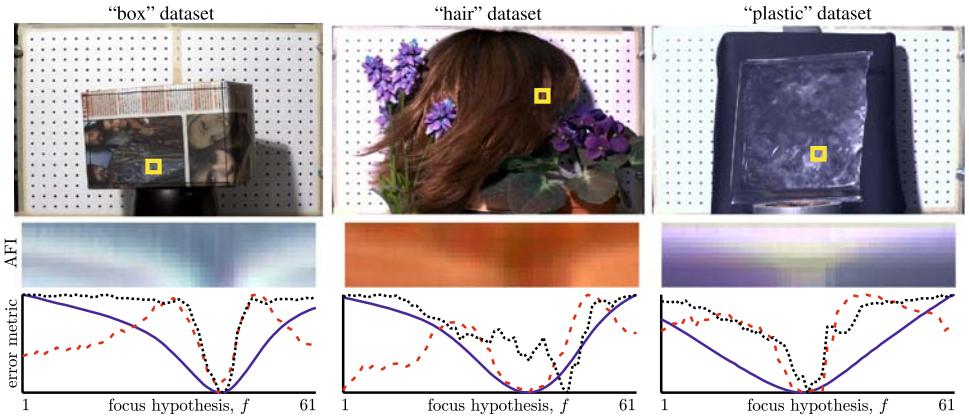


**Fig. 8** Behavior of focus criteria for a specific pixel (*highlighted square*) in three test datasets. The *dashed graph* is for direct confocal constancy (8), *solid* is for AFI model-fitting (9), and the *dotted graph* is for $3 \times 3$ variance (DFF). While all three criteria often have corresponding 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. 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

images respectively in RAW 12-bit mode. Each image was demosaiced using Canon software and linearized using the algorithm in Debevec and Malik (1997). We used $A = 13$ apertures ranging from f1.2 to f16, 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 6MP 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 $3072 \times 2048$. Unique to this setup, we manipulated focal setting using a computer-controlled stepping motor to drive the lens focusing ring mechanically (Technical Innovations). We used $A = 11$ apertures ranging from f3.5 to f16, 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 (Sect. 5) and then performed offline geometric calibration (Sect. 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. 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 Sect. 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 $128 \times 128$ pixel patch, a Matlab implementation of AFI model-fitting took about 250 seconds using $13 \times 61$ images, compared with 10 seconds for a depth from focus method that uses $1 \times 61$ images.

*Quantitative Evaluation: "Box" Dataset* To quantify reconstruction accuracy, we used a tilted planar scene consisting of a box wrapped in newsprint (Fig. 8, left). The plane of the box was measured using a FaroArm Gold 3D touch probe, as employed in Sect. 7.2, whose single-point accuracy was $\pm 0.05$ 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 (Bouguet 2004) 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 (8), AFI model-fitting (9), and a depth from focus (DFF) method, applied to

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

**Table 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 |
|---|---|---|---|---|
| $3 \times 3$ 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 |

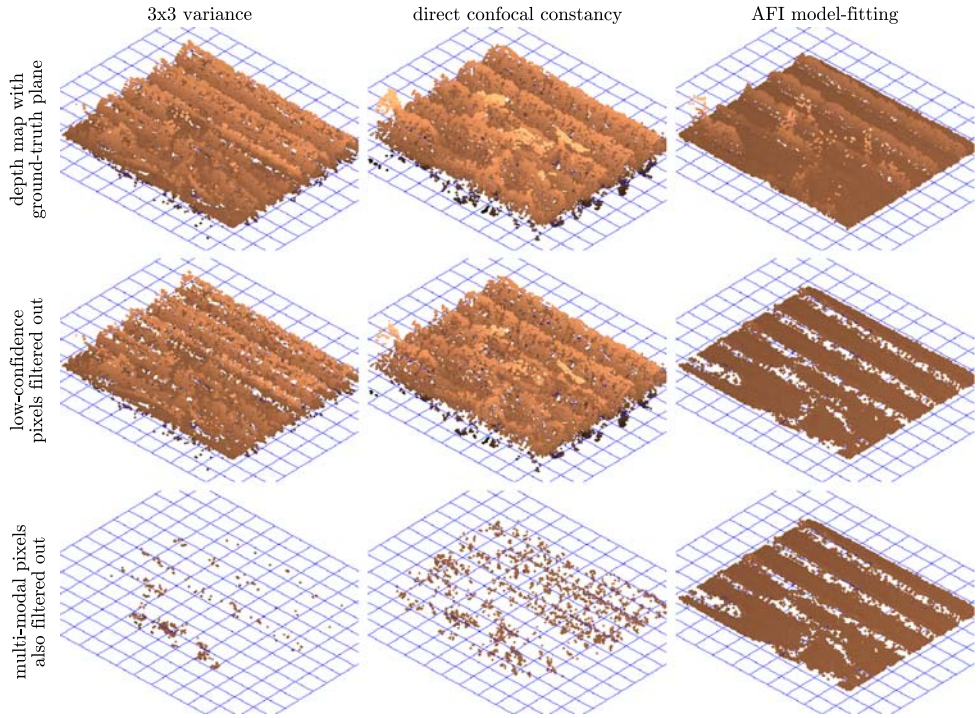


**Fig. 9** Visualizing the accuracy of reconstruction and outlier detection for the "box" dataset. *Top row:* For all three focus criteria, we show depth maps for a $200 \times 200$ region from the center of the box (see Fig. 10). The depth maps are rendered as 3D point clouds where intensity encodes 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

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 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 (Watanabe and Nayar 1998; Zhang and Nayar 2006), 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. 8) and for an image patch (Figs. 9 and 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

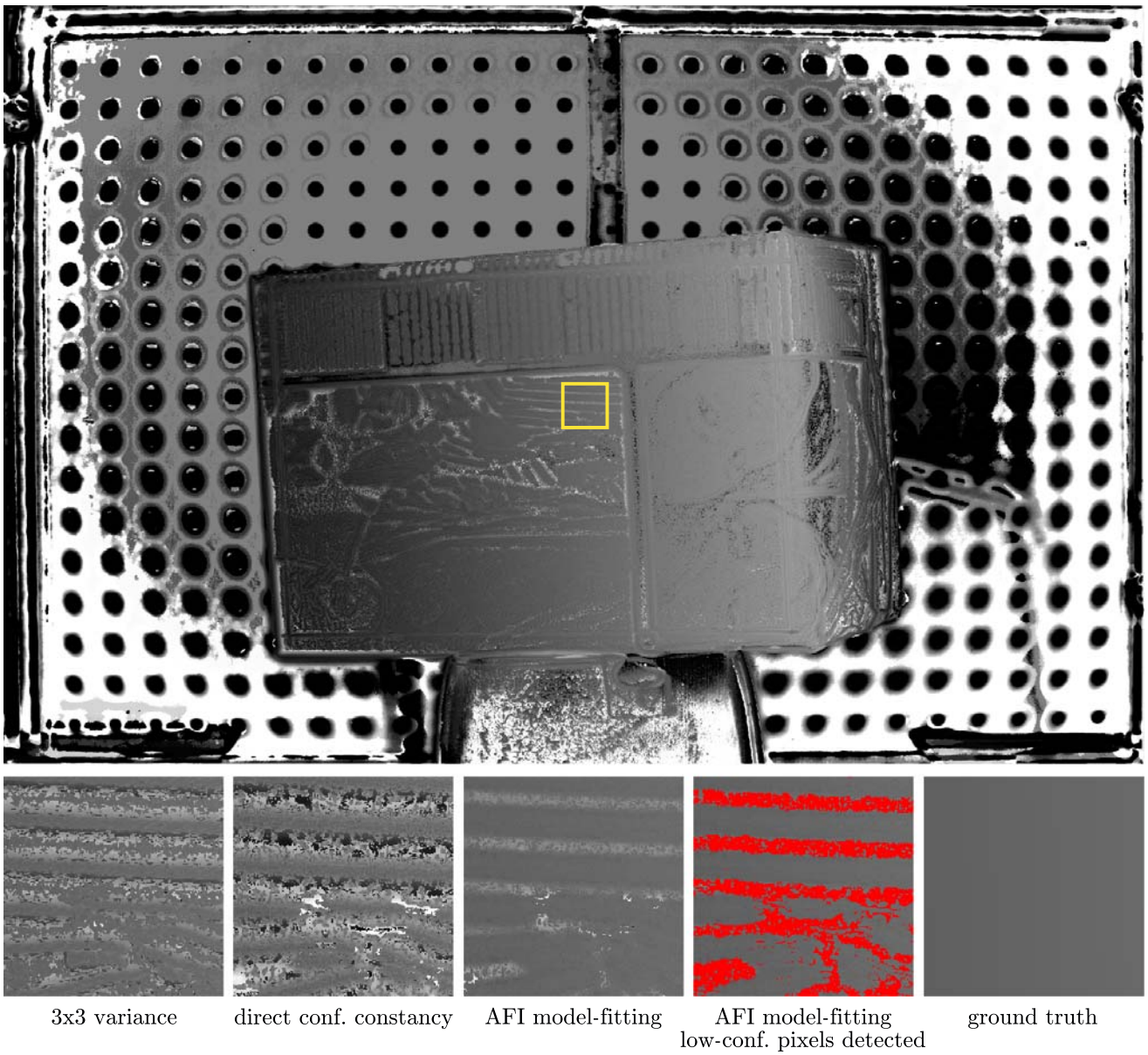


**Fig. 10** (Color online) *Top*: Depth map for the "box" dataset using AFI model-fitting. *Bottom*: Close-up depth maps for the highlighted region corresponding to Fig. 9, computed using three focus criteria

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. 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 reducing the apertures to five or six causes little reduction in reconstruction quality (Fig. 11).

*"Hair" Dataset* Our second test scene was a wig with a messy hairstyle, approximately 25 cm tall, surrounded by

[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.

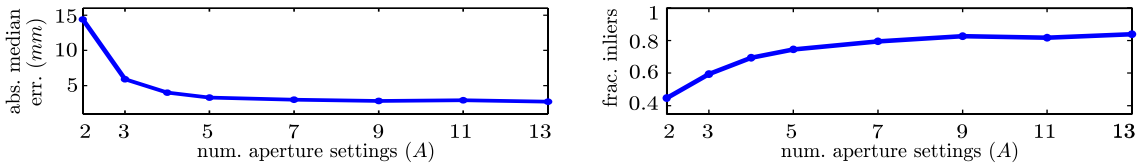


**Fig. 11** AFI model-fitting error and inlier fraction as a function of the number of aperture settings ("box" dataset, inlier threshold = 11 mm)

several artificial plants (Figs. 1 and 8, middle). Reconstruction results for this scene (Fig. 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 hair). By comparison, the $3\times3$ 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 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. 12 and 16). To better visualize the result of filtering out these pixels, we replace them using a simple variant of PDE-based inpainting (Bertalmio et al. 2000).

*"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. This plastic object was placed in front of a dark background and lit obliquely to enhance the contrast of its limited surface texture (Fig. 8, right). Reconstruction results for this scene (Figs. 13–14) 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. 14). By comparison, the $3\times3$ 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. 16).

*"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. 15). 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. 15, 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 (Sect. 6.4) as well as temporal variations in lighting intensity (Sect. 5), we compared the depth maps produced using AFI model-fitting (9) with and without this alignment step (Fig. 16a, 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.

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

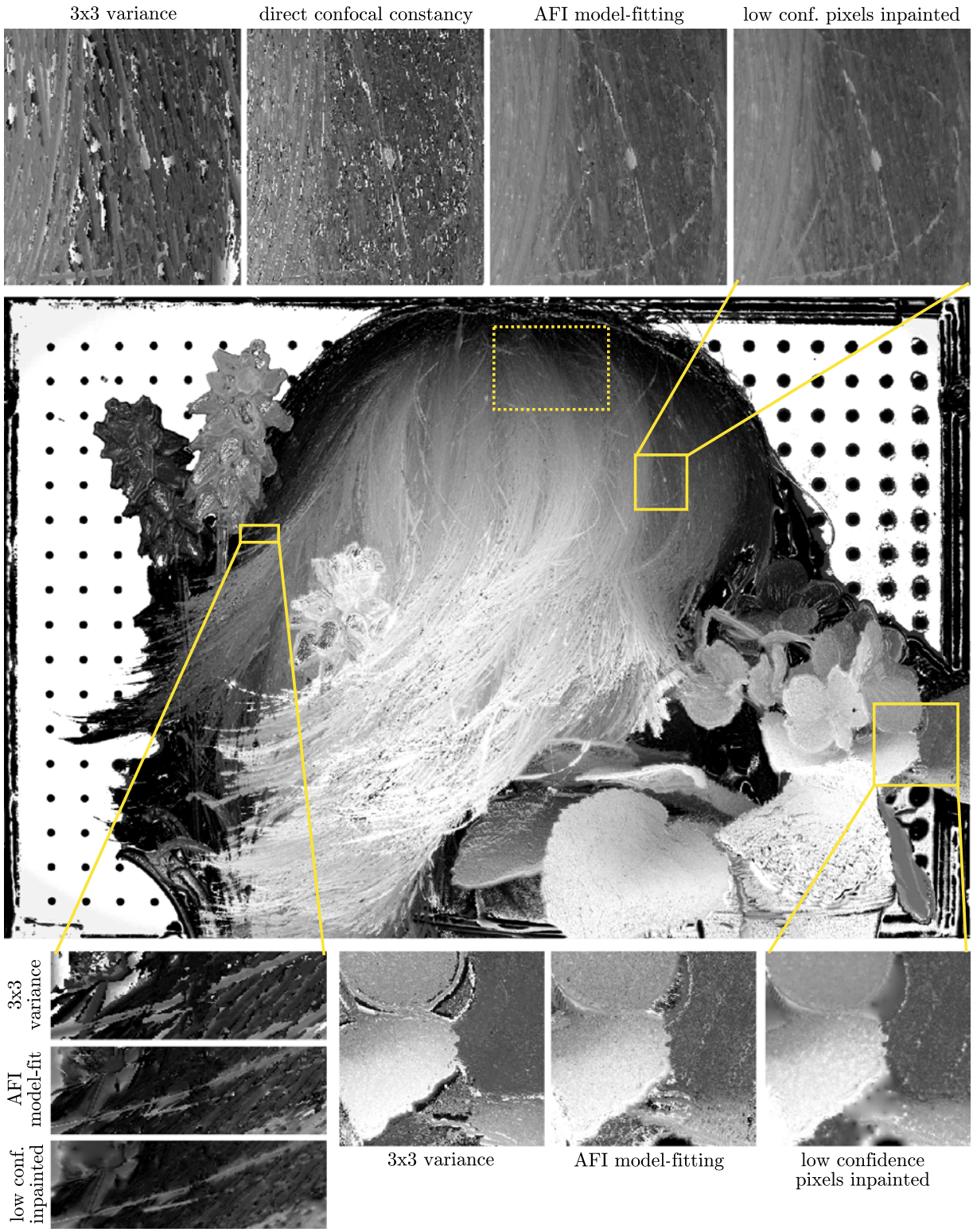


**Fig. 12** *Center*: Depth map for the "hair" dataset using AFI model-fitting. *Top*: The AFI-based depth map resolves several distinctive foreground strands of hair. We also show the result of detecting low-confidence pixels from AFI model-fitting and replacing them using PDE-based inpainting (Bertalmio et al. 2000) (see Fig. 16), 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, 3×3 variance (DFF) exhibits thick "halo" artifacts and fails to detect most of the foreground strands (see also Fig. 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

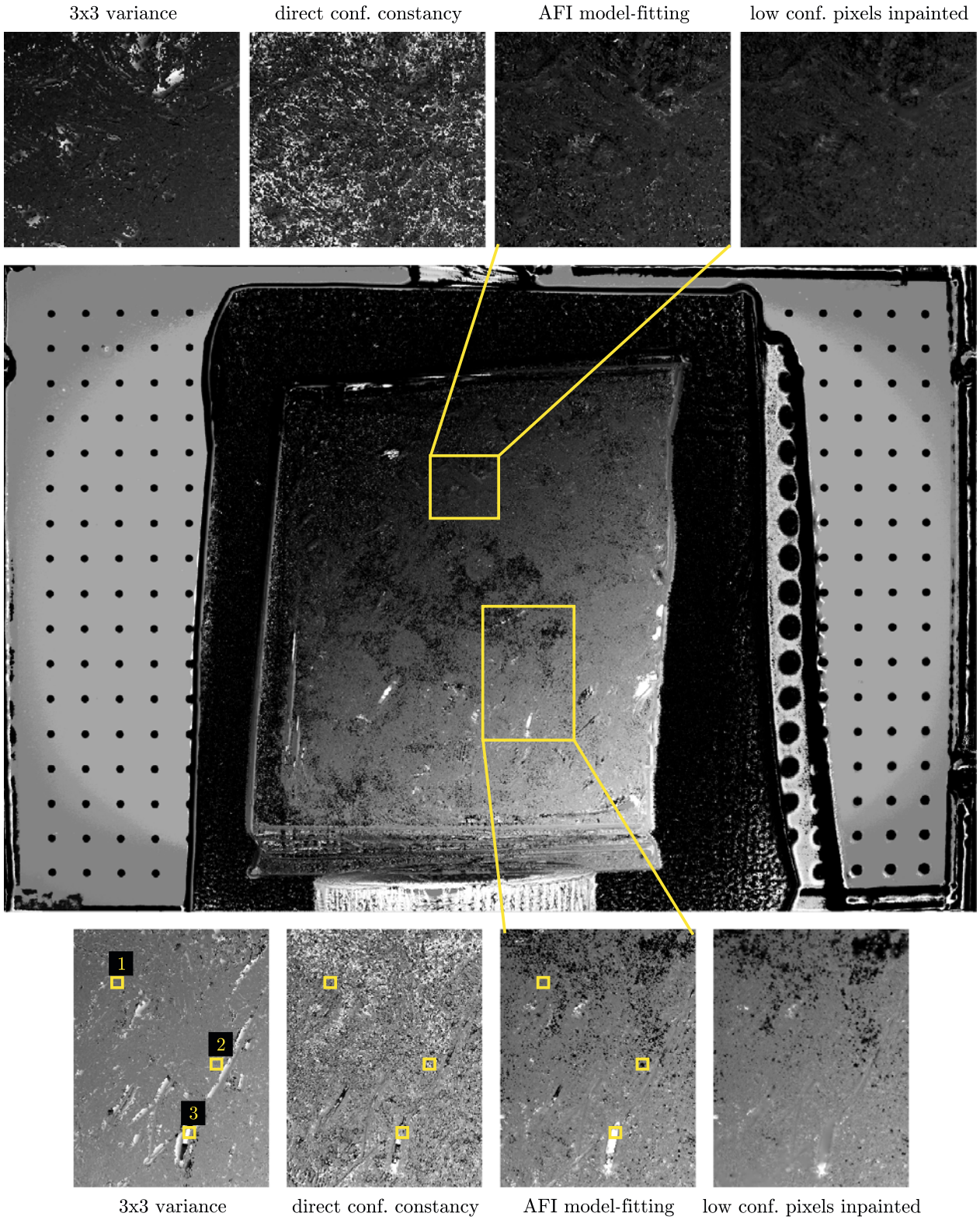


**Fig. 13** *Center*: Depth map for the "plastic" dataset using AFI model-fitting. *Top*: Close-up depth maps for the highlighted region, computed using three focus criteria. While $3 \times 3$ variance (DFF) yields the smoothest depth map overall for the transparent surface, there are still a significant 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. 14

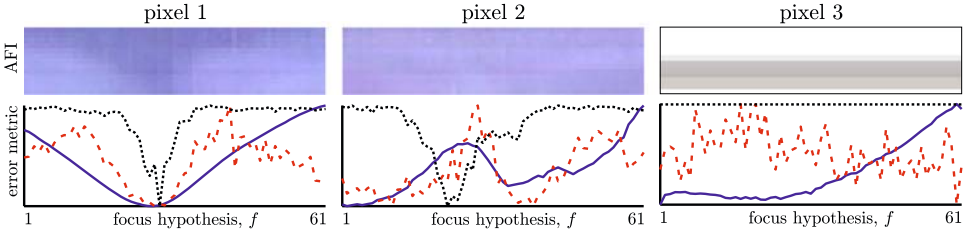


**Fig. 14** Failure examples. *Left to right*: Behavior of the three focus criteria in Fig. 13 for three highlighted pixels. The *dashed graph* is for direct confocal constancy (8), *solid* is for AFI model-fitting (9), and the *dotted graph* is for $3 \times 3$ variance (DFF). For pixel 1 all minima 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

points to provide useful depth estimates even in regions with relatively low texture.

## 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 (Favaro et al. 2003a; Vaish et al. 2006) or to reason about global scene geometry and occlusion (Asada et al. 1998b; Schechner and Kiryati 2000; Favaro and Soatto 2003).

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 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., (Zitnick et al. 2004), to the output of our per-pixel analysis, since (8) and (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.

[10]This follows from the work of Favaro et al. (2003a) who established that non-zero second-order albedo gradients are a necessary condition for resolving the structure of a smooth scene.

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. 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 boundaries 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 (Asada et al. 1998b; Favaro and Soatto 2003), as well as the use of a space-sweep approach to account for these occlusions, analogous to voxel-based stereo (Kutulakos and Seitz 2000).

## 10 Concluding Remarks

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.

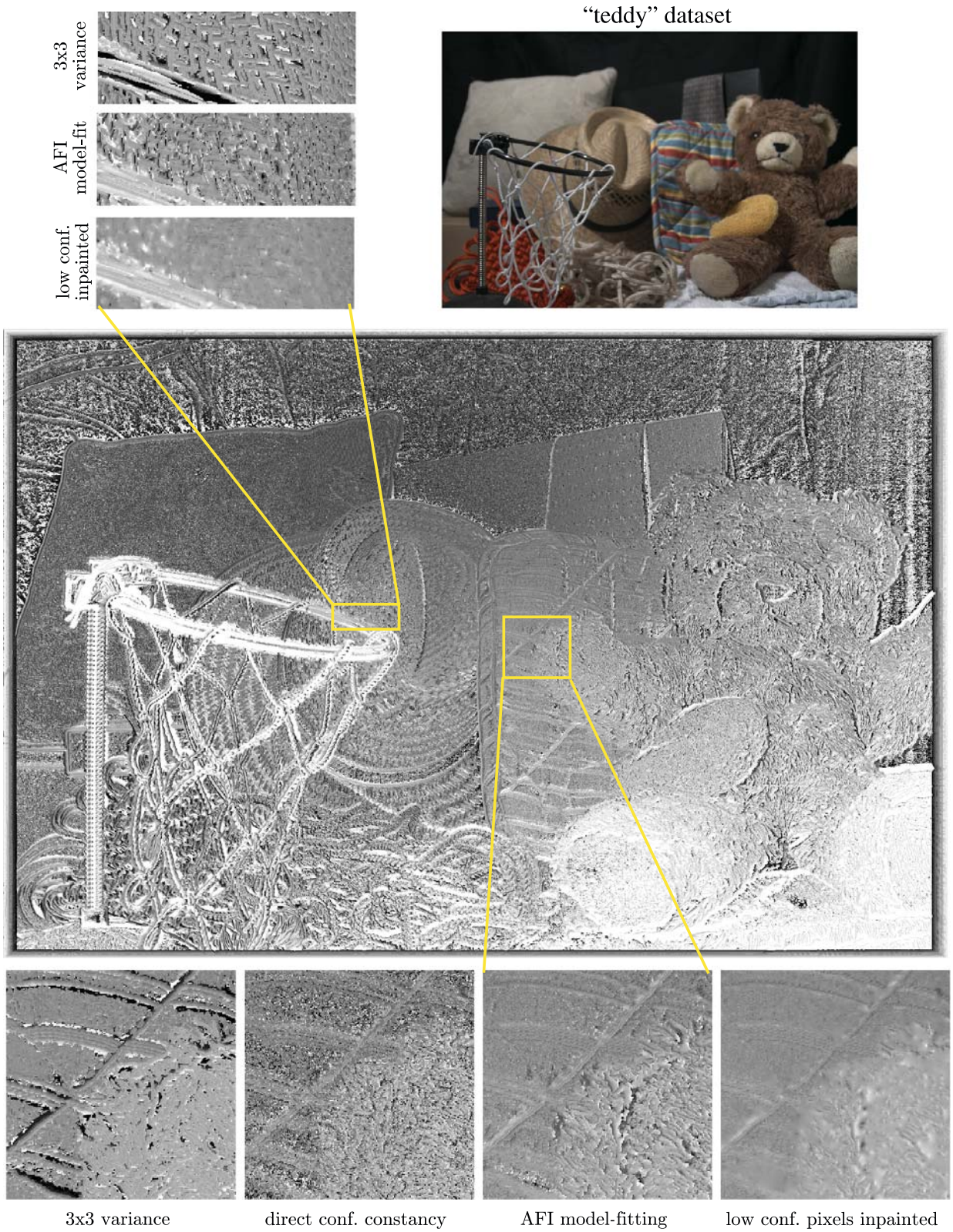


**Fig. 15** *Top right*: Sample widest-aperture f3.5 input photo of the "teddy" dataset. *Center*: Depth map using AFI model-fitting. *Top left:* Close-up depth maps for the highlighted region, comparing $3 \times 3$ variance (DFF) and AFI model-fitting, with and without inpainting of the detected outliers. Like the "plastic" dataset shown in Fig. 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

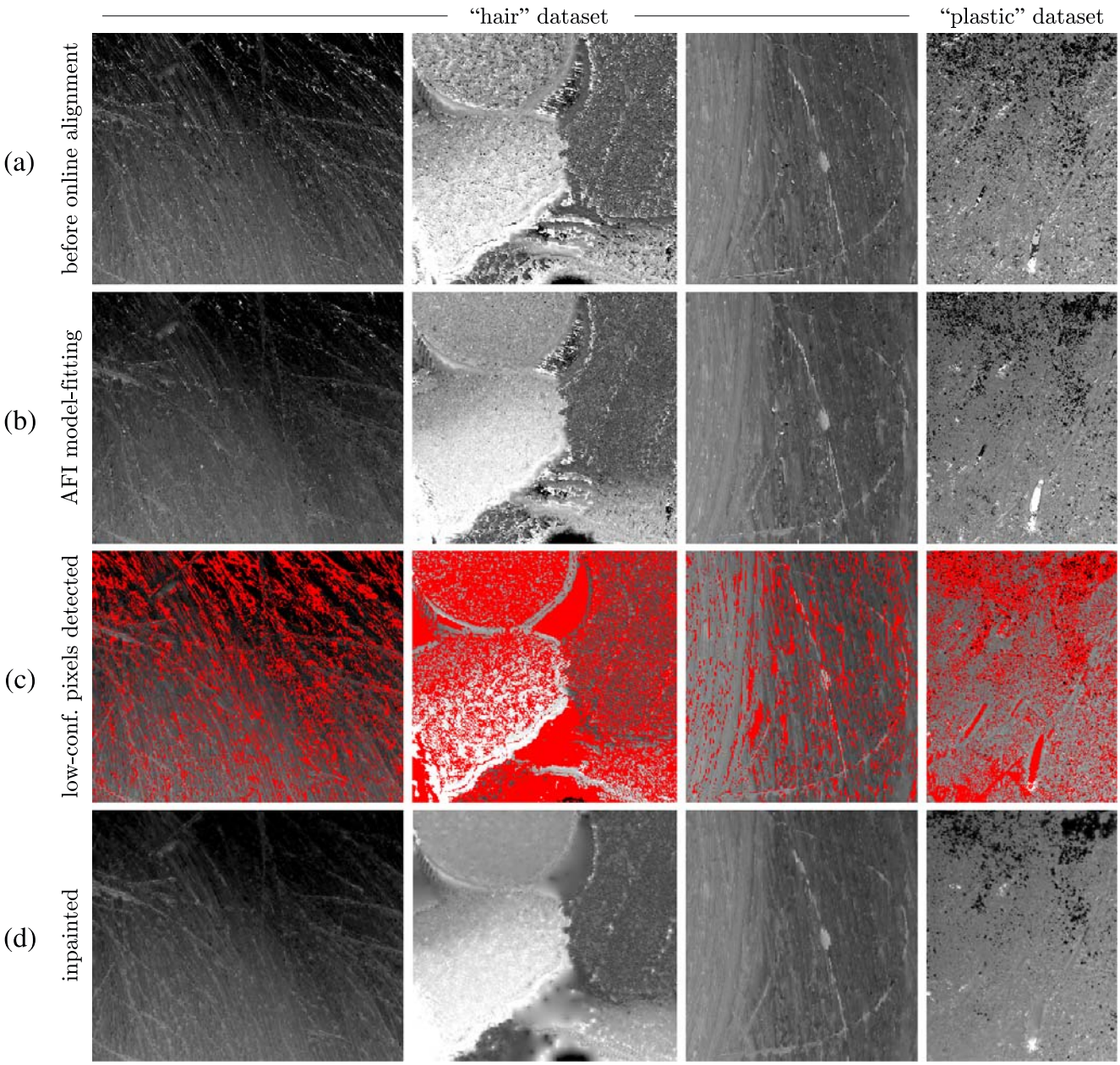


**Fig. 16** (Color online) (**a**)–(**b**) Improvement of AFI model-fitting due to online alignment, accounting for stochastic sub-pixel camera motion and temporal variations in lighting intensity. (**b**) Online alignment leads to a reduction in noisy pixels and yields smoother depth maps for low-textured 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. 9. (**d**) Low confidence pixels filled using PDE-based inpainting (Bertalmio et al. 2000). 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. 12–13

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 (Park et al. 2003), as well as for other applications.

**Acknowledgements** This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under the RG-PIN and CGS-D programs, by a fellowship from the Alfred P. Sloan Foundation, by an Ontario Premier's Research Excellence Award and by Microsoft Research.

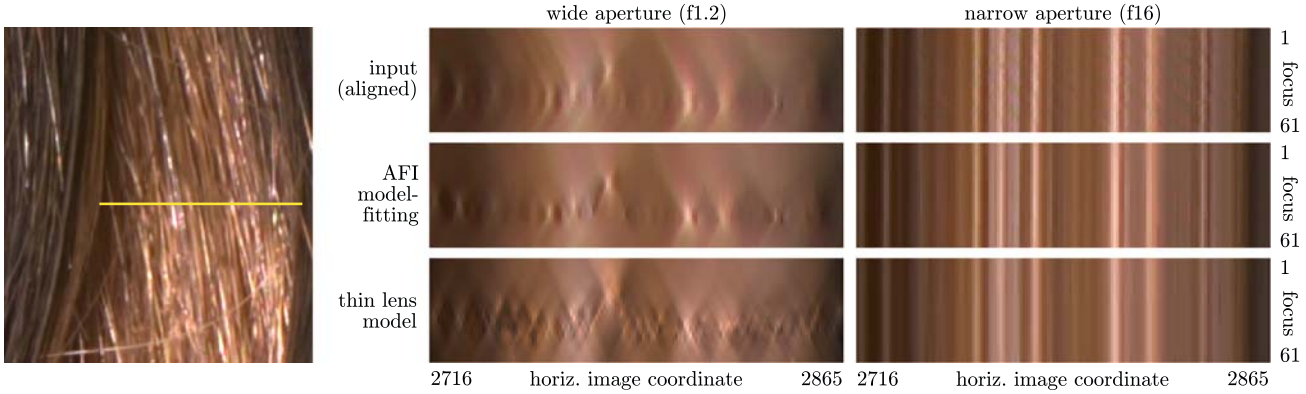


**Fig. 17** AFI model-fitting vs. the thin lens model. *Left*: Narrow-aperture image region from the "hair" dataset, corresponding to Fig. 12, top. *Right*: For two aperture settings, we show the cross-focus appearance variation of the highlighted horizontal segment: (i) for the aligned input images, (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

## Appendix A: Evaluation of Relative Exitance Recovery

To obtain a more quantitative evaluation of how well the relative exitance $R_{xy}(\alpha, f)$ can be recovered, and to validate that it does not depend on experimental conditions, we ran several additional experiments.

*Experiment 1* To test repeatability across different captures under fluorescent lighting, we repeated 5 trials of the radiometric calibration described in Sect. 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 Sect. 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* 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 are as follows:

| | Median mag. 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% |

Note that the median magnitude of the relative error corresponds to 1–2 gray levels out of 255.

*Experiment 3* We also compared the radiometric calibration from these experiments to the calibration used in Sect. 8, captured several years beforehand using 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* 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

Section 7.2 described how it is possible to approximate a pixel's AFI using a set of equi-blur regions where color

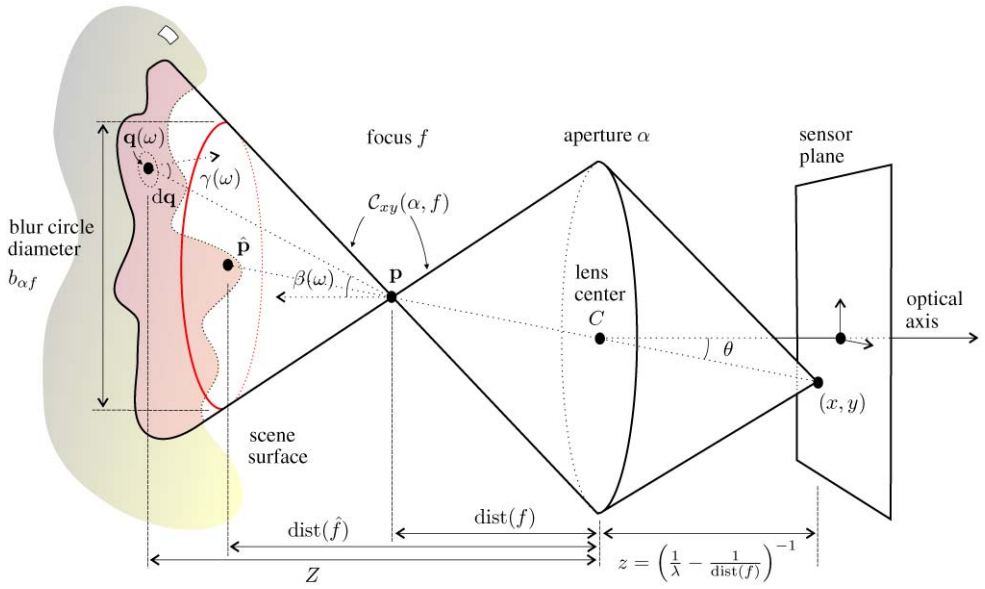


**Fig. 18** Thin lens imaging model for defocus (Asada et al. 1998b; Smith 2000). At an out-of-focus setting $f$, a point on the sensor plane $(x, y)$ integrates radiance from a region of the scene as shown. By contrast, at the perfect focus setting $\hat{f}$, all irradiance at $(x, y)$ would be due 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

and intensity remain constant. Here we establish Conditions 1–5 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. 18). We assume the following condition:

**Condition 1** Lens defocus can be described using the thin-lens model (Smith 2000; Asada et al. 1998b).

Then the image irradiance at $(x, y)$ is

$$E_{\alpha f}(x, y) = \frac{\pi\left(\frac{\lambda}{2\alpha}\right)^2 \cos^3\theta}{z^2} \times \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} \frac{L(\mathbf{q}(\omega), \omega) \cos\beta(\omega)}{\| \mathcal{C}_{xy}(\alpha, f) \|} \, \mathrm{d}\omega, \quad (11)$$

where $\frac{\lambda}{\alpha}$ is the aperture diameter; $\theta$ is the angle between the optical axis and the ray connecting $(x, y)$ and the lens center, $C$; $z = (\frac{1}{\lambda} - \frac{1}{\text{dist}(f)})^{-1}$ 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 (11) 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

$$b_{\alpha' f'} = b_{\alpha f} = \frac{\lambda}{\alpha} \frac{|\text{dist}(\hat{f}) - \text{dist}(f)|}{\text{dist}(f)}, \quad (12)$$

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:

**Condition 2** 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 (11), because it implies that $\beta(\omega) \to \theta$, giving

$$E_{\alpha f}(x, y) = \frac{\pi\left(\frac{\lambda}{2\alpha}\right)^2 \cos^4\theta}{z^2} \int_{\omega \in \mathcal{C}_{xy}(\alpha, f)} \frac{L(\mathbf{q}(\omega), \omega)}{\| \mathcal{C}_{xy}(\alpha, f) \|} \, \mathrm{d}\omega. \quad (13)$$

Note that the factor outside the integral in (13) is independent of the scene and accounted for by radiometric calibration (Sect. 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 (13) is simply the contribution to irradiance of a differential patch $\mathrm{d}\mathbf{q}$, centered on point $\mathbf{q}(\omega)$ and

subtending a solid angle of $d\omega$ from $\mathbf{p}$. Now consider the following two conditions:

**Condition 3** The outgoing radiance for any defocused scene point is constant within the cone subtended by the largest aperture, i.e., $L(\mathbf{q}(\omega), \omega) = L(\mathbf{q}(\omega))$.

**Condition 4** For any defocused scene point, the cone subtended by the largest aperture does not intersect the scene elsewhere.

Note that Conditions 3–4 are the same conditions required by confocal constancy (Sect. 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 (13), is independent of $(\alpha, f)$. Hence it suffices to show that the geometric factor $\frac{d\omega}{\|\mathcal{C}_{xy}(\alpha,f)\|}$ is independent of $(\alpha, f)$ in the same equi-blur region.

From the definition of solid angle, this factor is given by

$$\frac{d\omega}{\|\mathcal{C}_{xy}(\alpha, f)\|} = \frac{d\mathbf{q}\cos\gamma(\omega)\cos^2\beta(\omega)}{(Z - \mathrm{dist}(f))^2} \cdot \frac{\mathrm{dist}(f)^2}{\pi\left(\frac{\lambda}{2\alpha}\right)^2\cos^3\theta}, \tag{14}$$

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 $d\mathbf{q}$ and the ray connecting $\mathbf{q}(\omega)$ to $\mathbf{p}$.

Now assume that the following condition also holds:

**Condition 5** Depth variations for points within the defocused region of the scene approach zero, i.e., $Z \to \mathrm{dist}(\hat{f})$.

This condition implies that the depth, $Z$, of the differential patch $d\mathbf{q}$ can be approximated by the distance to the scene point $\hat{\mathbf{p}}$. We thus take $Z = \mathrm{dist}(\hat{f})$ and substitute (12) into (14), giving us a simplified version of (13):

$$E_{\alpha f}(x, y) = \frac{\left(\frac{\lambda}{\alpha}\right)^2\cos\theta}{z^2 b_{\alpha f}^2}\int_{\mathbf{q}\in\mathcal{C}_{xy}(\alpha,f)} L(\mathbf{q}(\omega), \omega)\cos^2\beta(\omega)\cos\gamma(\omega)d\mathbf{q}, \tag{15}$$

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 2, both $\beta(\omega)$ and $\gamma(\omega)$ will be constant over all $(\alpha, f)$. Therefore, the contribution of a differential scene patch $d\mathbf{q}$ 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 (15), 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 5, 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.

## References


Adelson, E. H., & Wang, J. Y. A. (1992). Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 99–106.

Asada, N., Fujiwara, H., & Matsuyama, T. (1998a). Edge and depth from focus. *International Journal of Computer Vision*, *26*(2), 153–163.

Asada, N., Fujiwara, H., & Matsuyama, T. (1998b). Seeing behind the scene: Analysis of photometric properties of occluding edges by the reversed projection blurring model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(2), 155–167.

Baker, S., & Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, *56*(3), 221–225.

Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. In *Proc. ACM SIGGRAPH* (pp. 417–424).

Bhasin, S. S., & Chaudhuri, S. (2001). Depth from defocus in presence of partial self occlusion. *Proc. International Conference on Computer Vision*, *2*, 488–493.

Bouguet, J.-Y. (2004). *Camera calibration toolbox for Matlab* (Oct. 14, 2004). http://vision.caltech.edu/bouguetj/calib_doc/.

Darrell, T., & Wohn, K. (1988). Pyramid based depth from focus. In *Proc. computer vision and pattern recognition* (pp. 504–509).

Debevec, P., & Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. In *Proc. ACM SIGGRAPH* (pp. 369–378).

Farid, H., & Simoncelli, E. P. (1998). Range estimation by optical differentiation. *Journal of the Optical Society of America A*, *15*(7), 1777–1786.

Favaro, P., & Soatto, S. (2002). Learning shape from defocus. *Proc. European Conference on Computer Vision*, *2*, 735–745.

Favaro, P., & Soatto, S. (2003). Seeing beyond occlusions (and other marvels of a finite lens aperture). In. *Proc. Computer Vision and Pattern Recognition*, *2*, 579–586.

Favaro, P., & Soatto, S. (2005). A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3).

Favaro, P., Mennucci, A., & Soatto, S. (2003a). Observing shape from defocused images. *International Journal of Computer Vision*, *52*(1), 25–43.

Favaro, P., Osher, S., Soatto, S., & Vese, L. A. (2003b). 3D shape from anisotropic diffusion. *Proc. Computer Vision and Pattern Recognition*, *1*, 179–186.

Fitzgibbon, A., Wexler, Y., & Zisserman, A. (2005). Image-based rendering using image-based priors. *International Journal of Computer Vision*, *63*(2), 141–151.

Fraser, C. S., & Shortis, M. R. (1992). Variation of distortion within the photographic field. *Photogrammetric Engineering and Remote Sensing*, *58*(6), 851–855.

Green, P., Sun, W., Matusik, W., & Durand, F. (2007). Multi-aperture photography. In *Proc. ACM SIGGRAPH*.

Grossberg, M. D., & Nayar, S. K. (2004). Modeling the space of camera response functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(10), 1272–1282.

Hasinoff, S. W., & Kutulakos, K. N. (2007). A layer-based restoration framework for variable-aperture photography. In *Proc. international conference on computer vision*.

Healey, G. E., & Kondepudy, R. (1994). Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(3), 267–276.

Hertzmann, A., & Seitz, S. M. (2005). Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1254–1264.

Isaksen, A., McMillan, L., & Gortler, S. J. (2000). Dynamically reparameterized light fields. In *Proc. ACM SIGGRAPH* (pp. 297–306).

Jin, H., & Favaro, P. (2002). A variational approach to shape from defocus. *Proc. European Conference on Computer Vision*, *2*, 18–30.

Kang, S. B., & Weiss, R. S. (2000). Can we calibrate a camera using an image of a flat, textureless Lambertian surface? *Proc. European Conference on Computer Vision*, *2*, 640–653.

Krotkov, E. (1987). Focusing. *International Journal of Computer Vision*, *1*(3), 223–237.

Kubota, A., Takahashi, K., Aizawa, K., & Chen, T. (2004). All-focused light field rendering. In *Proc. eurographics symposium on rendering*.

Kutulakos, K. N., & Seitz, S. M. (2000). A theory of shape by shape carving. *International Journal of Computer Vision*, *38*(3), 197–216.

Levin, A., Fergus, R., Durand, F., & Freeman, W. T. (2007). Image and depth from a conventional camera with a coded aperture. In *Proc. ACM SIGGRAPH*.

Levoy, M., & Hanrahan, P. (1996). Light field rendering. In *Proc. ACM SIGGRAPH* (pp. 31–42).

Levoy, M., Chen, B., Vaish, V., Horowitz, M., McDowall, I., & Bolas, M. T. (2004). Synthetic aperture confocal imaging. In *Proc. ACM SIGGRAPH* (pp. 825–834).

McGuire, M., Matusik, W., Pfister, H., Hughes, J. F., & Durand, F. (2005). Defocus video matting. In *Proc. ACM SIGGRAPH* (pp. 567–576).

Moreno-Noguer, F., Belhumeur, P. N., & Nayar, S. K. (2007). Active refocusing of images and videos. In *Proc. ACM SIGGRAPH*.

Nair, H., & Stewart, C. (1992). Robust focus ranging. In *Proc. computer vision and pattern recognition* (pp. 309–314).

Nayar, S., Watanabe, M., & Noguchi, M. (1996). Real-time focus range sensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(12), 1186–1198.

Ng, R. (2005). Fourier slice photography. In *Proc. ACM SIGGRAPH* (pp. 735–744).

Paris, S., Briceño, H., & Sillion, F. (2004). Capture of hair geometry from multiple images. In *Proc. ACM SIGGRAPH* (pp. 712–719).

Park, S. C., Park, M. K., & Kang, M. G. (2003). Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, *20*(3), 21–36.

Pentland, A. P. (1987). A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*(4), 523–531.

Rajagopalan, A. N., & Chaudhuri, S. (1999). An MRF model-based approach to simultaneous recovery of depth and restoration from defocused images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(7), 577–589.

Schechner, Y. Y., & Kiryati, N. (2000). Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, *39*(2), 141–162.

Smith, W. J. (2000). *Modern Optical Engineering* (3rd ed.) New York: McGraw-Hill.

Subbarao, M., & Surya, G. (1994). Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, *13*(3), 271–294.

Technical Innovations. http://www.robofocus.com/.

Vaish, V., Szeliski, R., Zitnick, C. L., & Kang, S. B. (2006). Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proc. computer vision and pattern recognition* (pp. 2331–2338).

Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., & Tumblin, J. (2007). Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *Proc. ACM SIGGRAPH*.

Watanabe, M., & Nayar, S. K. (1997). Telecentric optics for focus analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(12), 1360–1365.

Watanabe, M., & Nayar, S. K. (1998). Rational filters for passive depth from defocus. *International Journal of Computer Vision*, *27*(3), 203–225.

Webb, R. H. (1996). Confocal optical microscopy. *Reports on Progress in Physics*, *59*(3), 427–471.

Wei, Y., Ofek, E., Quan, L., & Shum, H.-Y. (2005). Modeling hair from multiple views. In *Proc. ACM SIGGRAPH* (pp. 816–820).

Willson, R. (1994a). Modeling and calibration of automated zoom lenses. In *Proc. SPIE #2350: Videometrics III* (pp. 170–186).

Willson, R. (1994b). *Modeling and calibration of automated zoom lenses*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Willson, R., & Shafer, S. (1994). What is the center of the image? *Journal of the Optical Society of America A*, *11*(11), 2946–2955.

Xiong, Y., & Shafer, S. (1997). Moment and hypergeometric filters for high precision computation of focus, stereo and optical flow. *International Journal of Computer Vision*, *22*(1), 25–59.

Zhang, L., & Nayar, S. K. (2006). Projection defocus analysis for scene capture and image display. In *Proc. ACM SIGGRAPH* (pp. 907–915).

Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S., & Szeliski, R. (2004). High-quality video view interpolation using a layered representation. In *Proc. ACM SIGGRAPH* (pp. 600–608).