A Layer-Based Restoration Framework for Variable-Aperture Photography

Samuel W. Hasinoff Kiriakos N. Kutulakos

University of Toronto
{hasinoff,kyros}@cs.toronto.edu

Abstract

We present variable-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.*

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 [16, 10, 2], or using specialized optics [17, 13]. 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 [16].

In this work we show that flexibility can be greatly increased through variable-aperture photography, *i.e.*, by collecting several images of the scene with all settings except aperture fixed (Figure 1). In particular, our method is designed to work with very few input images, including the three-image "aperture bracketing" mode found on consumer digital SLR cameras. In contrast to how easily one can obtain variable-aperture input images, controlling focus in a calibrated way requires special equipment on cur-

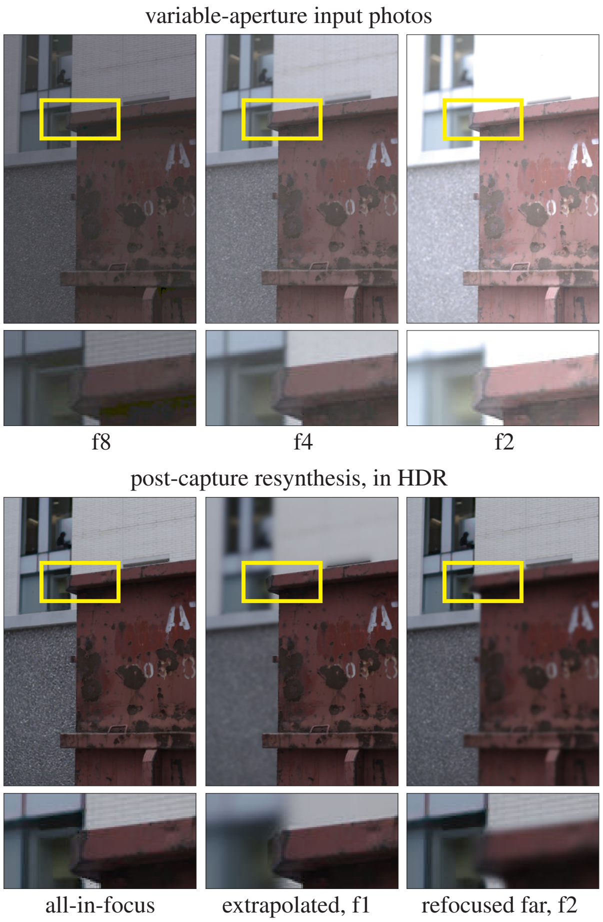

variable-aperture input photos
f8
f4
f2
post-capture resynthesis, in HDR
all-in-focus
extrapolated, f1
refocused far, f2

Figure 1. Variable-aperture photography. *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 (f1), and refocusing on the background (f2). See [1] for videos.

rent cameras. Variable-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 [5] in a physically meaningful way. Though several depth-from-defocus methods have previously addressed such occlusion, these methods have been limited by computational inefficiency [11], a restrictive occlusion model [7], or the assumption that the scene is composed of two surfaces [7, 11, 15]. 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*. [15], we formulate our image formation model in terms of image compositing [20], 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, that recover an all-in-focus representation of the scene [19, 14, 11, 21]. 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 [19, 14, 21] which causes serious 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 [4, 9], and do not model occlusion either. Importantly, none of these methods are designed to handle the large exposure differences found in variable-aperture photography.

Our work has four main contributions. First, we introduce variable-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, variable-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.

2. Variable-aperture photography

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 $\mathbf{R}$, the mean scene radiance $\overline{\mathbf{L}}$, the sensor response function $g(\cdot)$, and image noise η,

$$\mathbf{I}(x,y,a) = g\Big(\overbrace{\underbrace{\mathbf{R}(x,y,a,f)}_{\text{lens term}} \cdot \underbrace{\overline{\mathbf{L}}(x,y,a,f)}_{\text{scene radiance term}}}^{\text{sensor irradiance}}\Big) + \underbrace{\eta}_{\text{noise}}\ , \tag{1}$$

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 [16], 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. (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.

3. Image formation model

Sensor model. Following the high dynamic range literature [16], we express the sensor response $g(\cdot)$ in Eq. (1) as a

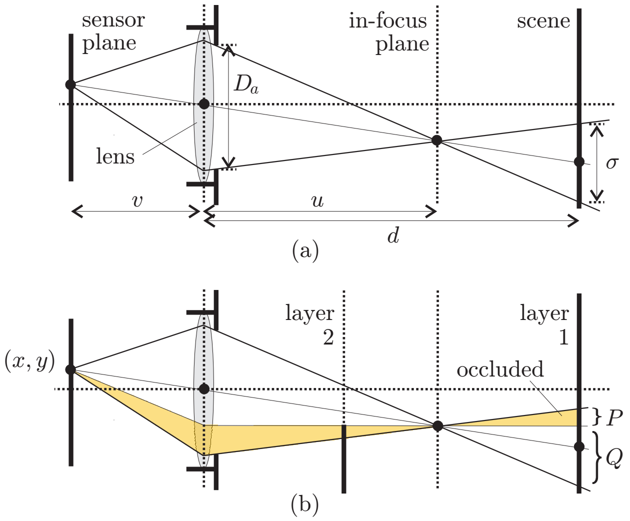


Figure 2. Defocused image formation with the thin lens model. (a) Fronto-parallel scene. (b) For a two-layered scene, the shaded fraction of the cone integrates radiance from layer 2 only, while the unshaded fraction integrates the unoccluded part of layer 1. Our occlusion model of Section 4 approximates layer 1's contribution to the radiance at (x, y) as $(\mathbf{L}_P + \mathbf{L}_Q)\frac{|Q|}{|P|+|Q|}$, which is a good approximation when $\frac{1}{|P|}\mathbf{L}_P \approx \frac{1}{|Q|}\mathbf{L}_Q$.

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 η, which we model as additive.

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, to represent the total solid angle subtended by the aperture, we use 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,

$$\mathbf{R}(x, y, a, f) = e_a\, \hat{\mathbf{R}}(x, y, a, f) \ , \tag{2}$$

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 [3], we assume that the defocus induced by the aperture obeys the standard thin lens model [18, 5]. 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 $\overline{\mathbf{L}} = \mathbf{L} * B_\sigma$ [18]. The 2D point-spread function B is parameterized by the effective *blur diameter*, σ, which depends on scene depth, focus setting, and aperture size (Figure 2a). From simple geometry,

$$\sigma = \frac{|d - u|}{u} D_a \ , \tag{3}$$

where d is the depth of the scene, u is the depth of the in-focus plane, and D_a is the diameter of the aperture. This implies that regardless of the scene depth, the blur diameter is proportional to the aperture diameter.

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_σ is a 2D symmetric Gaussian, where σ represents the standard deviation.

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, 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.[1] A fully accurate simulation of the thin lens model under occlusion involves backprojecting a cone into the scene, and integrating the unoccluded radiance (Figure 2b) [5]. Unfortunately, this process is computationally intensive, since the point-spread function can vary with arbitrary complexity according to the geometry of the occlusion boundaries.

To ensure tractability, we therefore formulate an approximate model for layered image formation (Figure 3) that accounts for occlusion, is designed to be efficiently computable and effective in practice, and leads to simple analytic gradients used for optimization.

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

$$\overline{L} = \sum_{k=1}^{K} [(\mathbf{A}_k \mathbf{L}_k) * B_{\sigma_k}] \cdot \mathbf{M}_k \ . \tag{4}$$

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

$$\mathbf{M}_k \ = \prod_{k'=k+1}^{K} \left(1 - \mathbf{A}_{k'} * B_{\sigma_{k'}}\right) \ . \tag{5}$$

[1] Since we model the layers as thin, occlusion due to perpendicular step edges [7] can be ignored.

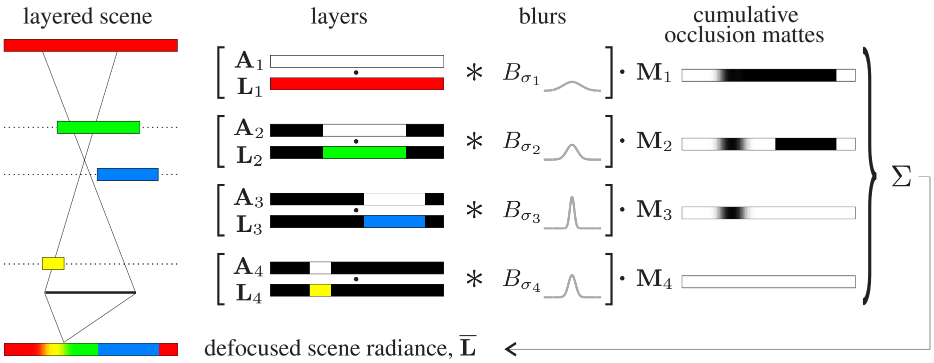

layered scene
layers
blurs
cumulative occlusion mattes
$\mathbf{A}_1$ $\mathbf{L}_1$ $*$ B_{σ_1} $\cdot$ $\mathbf{M}_1$
$\mathbf{A}_2$ $\mathbf{L}_2$ $*$ B_{σ_2} $\cdot$ $\mathbf{M}_2$
$\mathbf{A}_3$ $\mathbf{L}_3$ $*$ B_{σ_3} $\cdot$ $\mathbf{M}_3$
$\mathbf{A}_4$ $\mathbf{L}_4$ $*$ B_{σ_4} $\cdot$ $\mathbf{M}_4$
Σ
defocused scene radiance, $\overline{\mathbf{L}}$

Figure 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, $\overline{\mathbf{L}}$, we use convolution to independently defocus each layer $\mathbf{A}_k\mathbf{L}_k$, where the blur diameters σ_k are defined by the depths of the layers (Eq. (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.

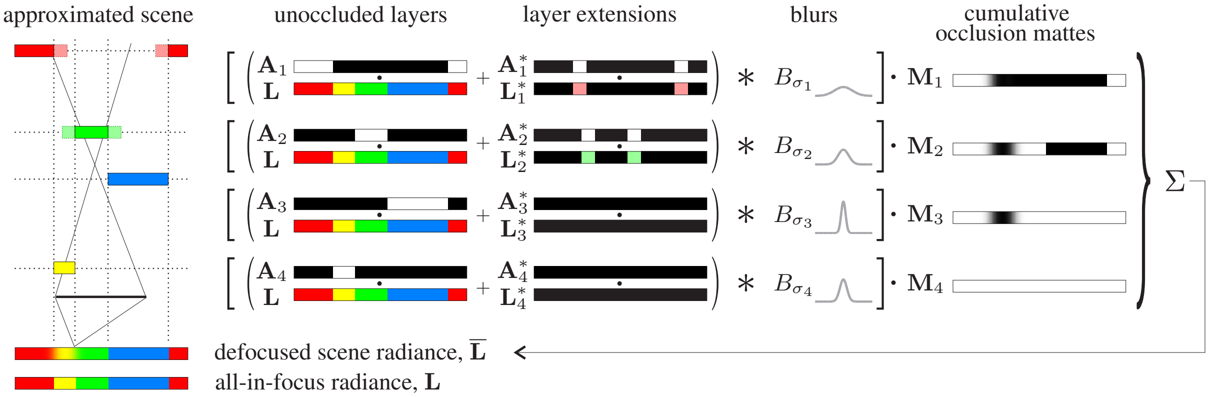

approximated scene
unoccluded layers
layer extensions
blurs
cumulative occlusion mattes
$\mathbf{A}_1$ $\mathbf{L}$ $+$ $\mathbf{A}_1^*$ $\mathbf{L}_1^*$ $*$ B_{σ_1} $\cdot$ $\mathbf{M}_1$
$\mathbf{A}_2$ $\mathbf{L}$ $+$ $\mathbf{A}_2^*$ $\mathbf{L}_2^*$ $*$ B_{σ_2} $\cdot$ $\mathbf{M}_2$
$\mathbf{A}_3$ $\mathbf{L}$ $+$ $\mathbf{A}_3^*$ $\mathbf{L}_3^*$ $*$ B_{σ_3} $\cdot$ $\mathbf{M}_3$
$\mathbf{A}_4$ $\mathbf{L}$ $+$ $\mathbf{A}_4^*$ $\mathbf{L}_4^*$ $*$ B_{σ_4} $\cdot$ $\mathbf{M}_4$
Σ
defocused scene radiance, $\overline{\mathbf{L}}$
all-in-focus radiance, $\mathbf{L}$

Figure 4. Reduced representation for the layered scene in Figure 3, based on the all-in-focus radiance, $\mathbf{L}$. The all-in-focus radiance specifies the unoccluded regions of each layer, $\mathbf{A}_k\mathbf{L}$, 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, $\mathbf{A}_k\mathbf{L} + \mathbf{A}_k^*\mathbf{L}_k^*$, we apply the same image formation model as in Figure 3.

Eqs. (4) and (5) can be viewed as an application of the matting equation [20], and generalizes the method of McGuire, *et al.* [15] to arbitrary focus settings and numbers of layers.

Intuitively, rather than integrating partial cones of rays that are restricted by the geometry of the occlusion boundaries (Figure 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 (Figure 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. In this representation, each layer is modeled as a binary alpha matte that "selects" the pixels of each layer (Figure 4).

While the all-in-focus radiance directly specifies the unoccluded radiance $\mathbf{A}_k\mathbf{L}$ for each layer, accurate modeling of defocus near occlusions requires an estimate of radiance at occluded points on the layers too (Figure 2b). We estimate extended versions of the unoccluded layers, $\mathbf{A}_k\mathbf{L} + \mathbf{A}_k^*\mathbf{L}_k^*$, in Section 7. The same image formation model of Eq. (4) applies in this case well.

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 segmentation of the scene into unoccluded layers, $\mathbf{A} = \{\mathbf{A}_k\}$, and the per-layer blur diameters, σ, specified in the widest aperture.[2]

[2] We use Eq. (3) to relate the blur diameters over aperture setting. In practice, however, we estimate the ratio of aperture diameters, D_a/D_A, using the calibrated exposure factors, *i.e.*, $\sqrt{e_a/e_A}$. This approach is more accurate than directly using the manufacturer-supplied f-numbers.

$$\mathbf{\Delta}(x,y,a) = \underbrace{\frac{1}{\hat{\mathbf{R}}(x,y,a,f)}\, g^{-1}\big(\mathbf{I}(x,y,a)\big)}_{\substack{\text{linearized and lens}-\text{corrected}\\ \text{image intensity}}} \;-\; \min\Bigg\{ \underbrace{e_a}_{\substack{\text{exposure}\\ \text{factor}}} \cdot \underbrace{\left[\sum_{k=1}^{K} \left[(\mathbf{A}_k \mathbf{L} + \mathbf{A}_k^{*} \mathbf{L}_k^{*}) * B_{\sigma_{a,k}}\right] \cdot \mathbf{M}_k\right]}_{\substack{\text{layered occlusion model}\\ \text{from Eqs. (4) and (5)}}},\; \underbrace{1}_{\substack{\text{clipping}\\ \text{term}}} \Bigg\}\,, \tag{7}$$

5. Restoration-based framework for HDR layer decomposition

In variable-aperture photography we do not have any prior information about 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 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

$$\mathcal{O}(\mathbf{L}, \mathbf{A}, \sigma) \;=\; \frac{1}{2} \sum_{a=1}^{A} \|\mathbf{\Delta}(x,y,a)\|^2 \;+\; \lambda\, \|\mathbf{L}\|_{\beta}\,. \tag{6}$$

In this optimization, $\mathbf{\Delta}(x,y,a)$ 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.

Eq. (7) shows the complete expression for the residual $\mathbf{\Delta}(x,y,a)$, parsed into simpler components. The residual is defined in terms of input images that have been linearized and lens-corrected. This transformation simplifies the optimization of Eq. (6), and converts the image formation model of Eq. (1) to scaling by an exposure factor e_a, followed by clipping to model over-saturation. Note that the transformation has the side-effect of amplifying the additive noise in Eq. (1),

$$\hat{\eta} = \frac{1}{\hat{\mathbf{R}}} \left| \frac{\mathrm{d}g^{-1}(\mathbf{I})}{\mathrm{d}\mathbf{I}} \right| \eta\,, \tag{8}$$

where $\hat{\eta} \;\rightarrow\; \infty$ for over-saturated pixels. Since this amplification can be quite significant, it must be taken into account during optimization. The innermost component of Eq. (7) is the layered image formation model of Section 4.

Weighted TV regularization. To regularize Eq. (6), we use a form of the total variation (TV) norm, $\|\mathbf{L}\|_{\mathrm{TV}} = \int \|\nabla \mathbf{L}\|$. This norm is useful for restoring sharp discontinuities, while suppressing noise and other high frequency detail [22]. The variant we propose,

$$\|\mathbf{L}\|_{\beta} = \int \sqrt{\big(w(\mathbf{L})\, \|\nabla \mathbf{L}\|\big)^2 + \beta}\,, \tag{9}$$

includes a perturbation term $\beta > 0$ that remains constant[3] and ensures differentiability as $\nabla \mathbf{L} \rightarrow 0$ [22]. 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 (Figure 7).

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 $\hat{\eta}$ predicted by Eq. (8).

6. Optimization method

To optimize Eq. (6), 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 [22, 21], by carrying out a set of conjugate gradient steps. Our formulation ensures that all required gradients have straightforward analytic formulas (Appendix A).
- **Blur refinement.** We use the same approach, of taking conjugate gradient steps, to optimize the blur diameters σ.
- **Layer refinement.** The layer decomposition $\mathbf{A}$ is more challenging to minimize because it involves a discrete labeling. 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 or behind the in-focus plane (Eq. (3)). We use a brute force approach that tests all 2^{K-1} distinct layer orderings, and select the one leading to the lowest objective (Figure 5c).
- **Initialization.** In order for this procedure to work, we need to initialize all three of $(\mathbf{L}, \mathbf{A}, \sigma)$, as discussed below.

7. Implementation details

Scene radiance initialization. We define an initial estimate for radiance, $\mathbf{L}$, by directly selecting pixels from the input images, scaled according to their exposure, e_a. For

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

each pixel, we choose the narrowest aperture for which the estimated signal-to-noise ratio, computed using Eq. (8), is above a fixed threshold. In this way, most pixels will come from the narrowest aperture image, except for the darkest regions of the scene, whose narrow-aperture pixel values will be dominated by noise.

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 [18, 9]. This method involves directly testing a set of hypotheses for blur diameter, specified in the widest aperture, by synthetically defocusing the image as if it were a fronto-parallel scene.

Because of the large exposure differences between photos taken several f-stops apart, we evaluate consistency with a given blur hypothesis by comparing images captured with successive aperture settings, $(a, a+1)$. To evaluate each such pair, we convolve the narrower aperture image with the incremental blur aligning it with the wider one. Since our point-spread function is Gaussian, this incremental blur can be expressed in a particularly simple form, namely another 2D Gaussian with standard deviation $(\sigma_{a+1}^2 - \sigma_a^2)^{\frac{1}{2}}$.

Each blur hypothesis therefore leads to a per-pixel error measuring how well the input images are resynthesized. We minimize this error within a Markov random field (MRF) framework, which allows us to reward global piecewise smoothness as well (Figure 5). In particular, we employ graph cuts with the expansion-move approach [8], where the smoothness cost is defined as a truncated linear function of adjacent label differences on the four-connected grid.

Sensor response and lens term calibration. To recover the sensor response function, $g(\cdot)$, we apply standard HDR imaging methods [16] to a calibration sequence captured with varying exposure time.

We recover the radiometric lens term $\mathbf{R}(x, y, a, f)$ using calibration as well, using the pixel-wise method in [12].

Occluded radiance estimation. As illustrated in Figure 4, we assume that all scene layers, even where occluded, can be expressed in terms of the all-in-focus radiance $\mathbf{L}$. In practice, we use inpainting to extend the unoccluded layers, by up to the largest blur diameter, behind any occluders. During optimization, we use a low-cost technique that simply chooses the nearest unoccluded pixel for a particular layer, but for rendering we use a higher-quality PDE-based inpainting method [6].

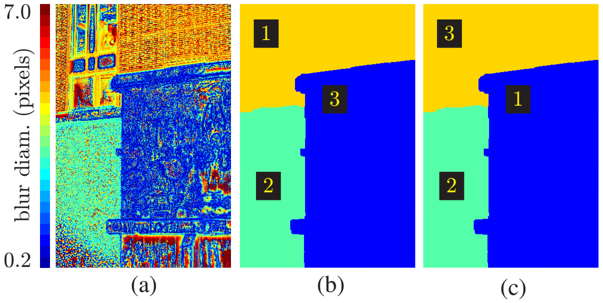


Figure 5. (a)–(b) Initial layer decomposition and blur assignment for the DUMPSTER dataset, obtained using our depth-from-defocus method: (a) greedy layer assignment, (b) MRF-based layer decomposition, with initial front-to-back depth ordering indicated. (c) Revised layering, obtained by iteratively modifying the layer assignment for high-residual pixels, and re-estimating the depth ordering.

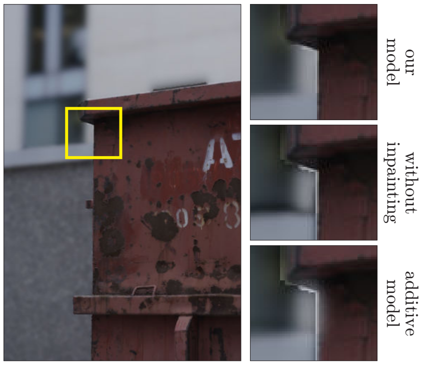


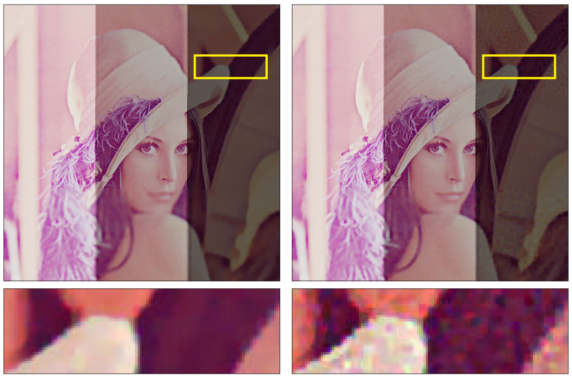
Figure 6. Layered image formation results at occlusion boundaries. *Left:* Tone-mapped HDR image of the DUMPSTER dataset, for an extrapolated aperture (f1). *Top inset:* Our model handles occlusions in a visually realistic way. *Middle:* Without inpainting, *i.e.*, assuming zero radiance in occluded regions, the resulting darkening emphasizes pixels whose layer assignment has been misestimated, that are not otherwise noticeable. *Bottom:* An additive image formation model [19, 21] exhibits similar artifacts, plus erroneous spill from the occluded background layer.

8. Results and discussion

To test our approach on real data, we captured sequences using a Canon EOS 1Ds Mark II, secured on a tripod, with an 85mm f1.2L lens set to manual focus. In all our experiments we use the three-image "aperture bracketing" mode set to ±2 stops, and select shutter speed so that the images are captured at f8, f4, and f2 (yielding relative exposure levels of roughly 1, 4, and 16, respectively). Adding more input images (*e.g.*, at half-stop intervals) does improve results, although less so in dark and defocused regions, which must be restored with deconvolution. We captured RAW images for increased dynamic range, and demonstrate our results for downsampled 500×333 pixel images.[4]

We also tested our approach using a synthetic dataset (LENA), to enable comparison with ground truth (Figure 7 and 8a). 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 decompose the image into layers by assigning different depths to each of three horizontal bands, and generate the input images by applying the forward im-

[4] See [1] for additional results and videos.



(a) (b)

Figure 7. Effect of TV weighting. 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, 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.

age formation model. Finally, we add Gaussian noise to the input with a standard deviation of 1% of the intensity range.

To obtain our results, we follow the iterative method described in Section 6, alternating 10 conjugate gradient steps each of image restoration and blur refinement, until convergence, interspersing the layer refinement and reordering procedure every 80 such steps. For all experiments we set the smoothing parameter to $\lambda = 0.002$.

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 (Figures 1, 6–8). Note that since we do not record the focus setting f at capture time, we only recover layer depths up to scale. Thus, to modify focus setting, we specify the depth of the in-focus plane as a fraction of the corresponding depth in the input. To help visualize the full exposure range of the HDR images, we apply tone-mapping using a simple global operator of the form $T(x) = \frac{x}{1+x}$.

For ease of comparison, we do not resynthesize the residual radiometric distortions $\hat{\mathbf{R}}$, such as vignetting, nor do we simulate geometric distortions, such as the image magnification caused by changing focus setting. If desired, these lens-specific artifacts can be simulated as well.

Note that while camera settings can also be extrapolated, this functionality is somewhat limited. In particular, while extrapolated wider apertures can model the increased relative defocus between layers (Figure 1, bottom), our input images lack the information needed to decompose an in-focus layer, wholly within the depth-of-field of the widest aperture, into any finer gradations of depth.

To evaluate our layered occlusion model in practice, we compare our resynthesis results at layer boundaries with those obtained using alternative methods. As shown in Figure 6, our layered occlusion model produces visually realistic output, and is a significant improvement over the additive model [19, 21]. Importantly, our layered occlusion model is accurate enough to resolve the correct layer ordering in all of our experiments, simply by applying brute force search, testing which ordering leads to the smallest objective.

Another strength of variable-aperture photography is that dark and defocused areas of the scene are handled naturally by our image restoration framework. These areas normally present a special challenge, since they are dominated by noise for narrow apertures, but defocused for wide apertures. In general, high-frequencies cannot be recovered in such regions, however, our variant of TV regularization helps successfully "deconvolve" blurred intensity edges and to suppress the effects of noise (Figure 7a, inset).

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 is sometimes not discriminative enough to identify pixels with incorrect layer labels, amidst 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 (Figures 6 and 8b). For further results and discussion of failure cases, see [1].

9. Concluding remarks

We demonstrated how variable-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. For future work, we are interested in extending our technique to multi-resolution, and addressing motion between exposures, possibly by incorporating optical flow into the optimization.

Acknowledgements This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under the RGPIN and CGS-D programs, by a fellowship from the Alfred P. Sloan Foundation, and by an Ontario Premiers Research Excellence Award.

A. Analytic gradient computation

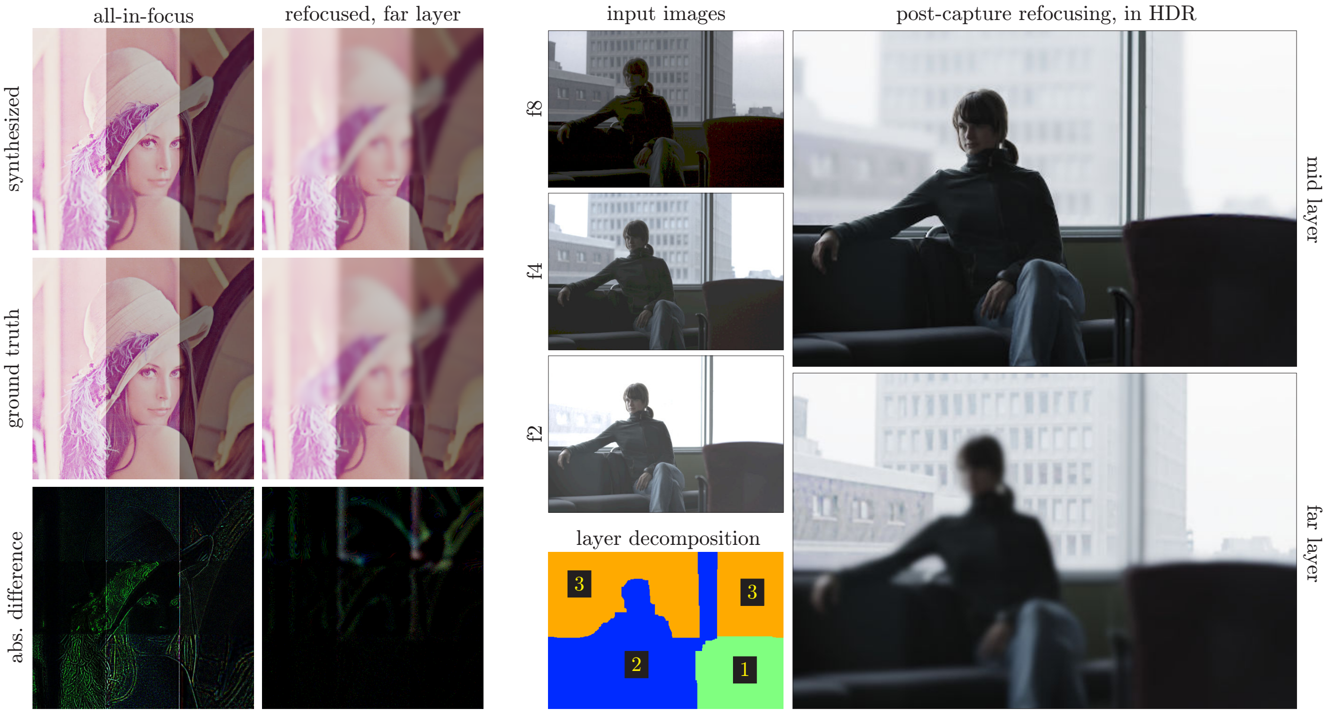
Because our image formation model is a simple linear operator, the gradients required to optimize our objective function take a compact analytic form.

Due to space constraints, the following expressions assume a single aperture only, with no inpainting (see the supplementary materials [1] for the generalization):

$$\frac{\partial \mathcal{O}}{\partial \mathbf{L}} = -\sum_{k=1}^{K} \left[\boldsymbol{\Delta} \mathbf{A}_k \mathbf{M}_k \star B_{\sigma_k} \right] + \frac{\partial \|\mathbf{L}\|_\beta}{\partial \mathbf{L}} \tag{10}$$

$$\frac{\partial \mathcal{O}}{\partial \sigma_k} = -\sum_{x,y} \left[\sum_{k'=1}^{K} \left[\boldsymbol{\Delta} \mathbf{A}_{k'} \mathbf{M}_{k'} \star \frac{\partial B_{\sigma_{k'}}}{\partial \sigma_{k'}} \right] \right] \mathbf{A}_k \mathbf{L} \, , \tag{11}$$

where $\star$ denotes 2D correlation, and these gradients are revised to be zero for over-saturated pixels. The gradient for



(a)

(b)

Figure 8. (a) Resynthesis results for the LENA dataset are almost visually indistinguishable ground truth, however slight differences, mainly due to image noise, remain. (b) For the PORTRAIT dataset, the gamma-corrected input images ($\gamma = 3$) show posterization artifacts because the scene's dynamic range is large. Although the final layer assignment has residual errors near boundaries, the restoration results are sufficient to resynthesize visually realistic new images. We demonstrate refocusing in HDR, simulating the widest input aperture (f2).

the regularization term is

$$\frac{\partial \|\mathbf{L}\|_\beta}{\partial \mathbf{L}} = -\operatorname{div}\left(\frac{w(\mathbf{L})^2\,\nabla\mathbf{L}}{\sqrt{\left(w(\mathbf{L})\,\|\nabla\mathbf{L}\|\right)^2+\beta}}\right) . \qquad (12)$$

References

[1] http://www.cs.toronto.edu/~hasinoff/aperture. 1, 6, 7

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *Proc. SIGGRAPH*, 23(3):294–302, 2004. 1

[3] M. Aggarwal and N. Ahuja. A pupil-centric model of image formation. *IJCV*, 48(3):195–214, 2002. 3

[4] K. Aizawa, K. Kodama, and A. Kubota. Producing object-based special effects by fusing multiple differently focused images. *TCSVT*, 10(2), 2000. 2

[5] N. Asada, H. Fujiwara, and T. Matsuyama. Seeing behind the scene: Analysis of photometric properties of occluding edges by the reversed projection blurring model. *TPAMI*, 20(2):155–167, 1998. 2, 3

[6] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. SIGGRAPH*, pp. 417–424, 2000. 6

[7] S. S. Bhasin and S. Chaudhuri. Depth from defocus in presence of partial self occlusion. In *Proc. ICCV*, vol. 2, pp. 488–493, 2001. 2, 3

[8] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 23(11):1222–1239, 2001. 6

[9] S. Chaudhuri. Defocus morphing in real aperture images. *JOSA A*, 22(11):2357–2365, 2005. 2, 6

[10] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.*, 23(3):673–678, 2004. 1

[11] P. Favaro and S. Soatto. Seeing beyond occlusions (and other marvels of a finite lens aperture). In *Proc. CVPR*, vol. 2, pp. 579–586, 2003. 2

[12] S. W. Hasinoff and K. N. Kutulakos. Confocal stereo. In *Proc. ECCV*, vol. 1, pp. 620–634, 2006. 6

[13] A. Isaksen, L. McMillan, and S. J. Gortler. Dynamically reparameterized light fields. In *Proc. SIGGRAPH*, pp. 297–306, 2000. 1

[14] H. Jin and P. Favaro. A variational approach to shape from defocus. In *Proc. ECCV*, vol. 2, pp. 18–30, 2002. 2

[15] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand. Defocus video matting. In *Proc. SIGGRAPH*, pp. 567–576, 2005. 2, 4

[16] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *Proc. CVPR*, pp. 1374–1380, 1999. 1, 2, 6

[17] R. Ng. Fourier slice photography. In *Proc. SIGGRAPH*, pp. 735–744, 2005. 1

[18] A. P. Pentland. A new sense for depth of field. *TPAMI*, 9(4):523–531, 1987. 3, 6

[19] A. N. Rajagopalan and S. Chaudhuri. An MRF model-based approach to simultaneous recovery of depth and restoration from defocused images. *TPAMI*, 21(7):577–589, 1999. 2, 6, 7

[20] A. Smith and J. Blinn. Blue screen matting. In *Proc. SIGGRAPH*, pp. 259–268, 1996. 2, 4

[21] M. Šorel and J. Flusser. Simultaneous recovery of scene structure and blind restoration of defocused images. In *Proc. Comp. Vision Winter Workshop*, pp. 40–45, 2006. 2, 5, 6, 7

[22] C. Vogel and M. Oman. Fast, robust total variation based reconstruction of noisy, blurred images. *TIP*, 7(6):813–824, 1998. 5