A Novel Depth from Defocus Framework Based on a Thick Lens Camera Model

Matthew Bailey Jean-Yves Guillemaut
Centre for Vision, Speech and Signal Processing
University of Surrey, UK
{m.j.bailey, j.guillemaut}@surrey.ac.uk

Abstract

Reconstruction approaches based on monocular defocus analysis such as Depth from Defocus (DFD) often utilise the thin lens camera model. Despite this widespread adoption, there are inherent limitations associated with it. Coupled with invalid parameterisation commonplace in literature, the overly-simplified image formation it describes leads to inaccurate defocus modelling; especially in macro-scale scenes. As a result, DFD reconstructions based around this model are not geometrically consistent, and are typically restricted to single-view applications. Subsequently, the handful of existing approaches which attempt to include additional viewpoints have had only limited success.

In this work, we address these issues by instead utilising a thick lens camera model, and propose a novel calibration procedure to accurately parameterise it. The effectiveness of our model and calibration is demonstrated with a novel DFD reconstruction framework. We achieve highly detailed, geometrically accurate and complete 3D models of real-world scenes from multi-view focal stacks. To our knowledge, this is the first time DFD has been successfully applied to complete scene modelling in this way.

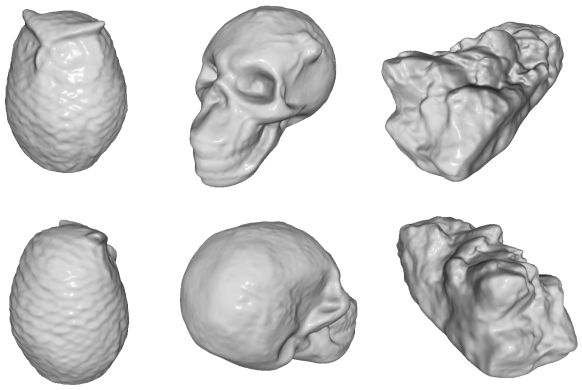


Figure 1. We propose a DFD pipeline that incorporates a thick lens camera model, which we parameterise using a novel calibration procedure. Multi-view focal stacks of macro-scale scenes are reconstructed using our iterative DFD framework, and combined to produce complete 3D models. To our knowledge, this is the first time DFD has been applied in this way.

1. Introduction

When reconstructing a scene from RGB images, it is commonplace to simplify the image formation process by making assumptions about the camera. For example, multi-view stereo (MVS) typically assumes a pinhole model [15], while focus-based approaches usually adopt a thin lens model. Although these models provide a mathematically convenient and often reasonable approximation, neither fully describe the behaviour of a modern lens.

Due to the ubiquious reliance on these simplified models, many reconstruction algorithms are implicitly restricted by the inaccuracies they introduce. In the context of DFD, the use of the thin lens model severely limits the accuracy of reconstruction in several ways.

First, although large focal stacks are benefical to reconstruction, most approaches do not consider focal stacks with more than two images. One reason is the depth ambiguities introduced by the thin lens model, which is exacerbated with the addition of more images. Other reasons include a lack of publically available datasets, as well as the absence of a standard approach for readily acquiring focal stacks.

Second, traditional DFD is generally limited to coarse, single-view reconstructions. Given the erroneous defocus modelling, depth resolution cannot be easily increased without heavy reliance on scene priors. As a result, general DFD approaches tend to produce low fidelity depth maps. Extending to multiple views is difficult because of the low quality, geometrically inconsistent reconstructions; as well as the inherent limitations of the thin lens model which we will discuss later.

While some works pursue approaches that iteratively adjust the defocus functions during the reconstruction to partially overcome these limitations [21, 8, 22, 23], we instead consider a different camera model to address these problems at their source. In this work, we adopt a thick lens

Published in: 2020 International Conference on 3D Vision (3DV), pages 1206–1215, 2020. DOI: https://doi.org/10.1109/3DV50981.2020.00131

model and apply it to multi-view DFD reconstruction (see Figure 1). To our knowledge, this has not been attempted previously due to the complexities this introduces during acquisition. However, the adoption of this model solves many of the discussed issues.

We present a novel and practical method for fully calibrating a camera as a thick lens that is applicable to the capture of multi-view focal stacks. Then, we apply our model to a novel DFD reconstruction framework, which allows us to demonstrate the improvements in scene modelling our approach has over traditional methods. In summary, this work makes the following key contributions:

1. An image formation model that better describes the appearance of images captured with a finite aperture
2. A practical thick lens calibration procedure for multi-view focal stacks
3. An MRF-based DFD reconstruction framework that produces high quality 3D reconstructions of macro-scale scenes

Our work is particuarly useful when applied to macro-scale scenes, since the limited depth of field (DoF) makes other approaches such as MVS unsuitable. Additionally, the monocular nature of DFD makes the reconstruction of complex materials that are traditionally challenging for MVS possible. Finally, few viewpoints are needed for complete scene coverage compared to conventional approaches.

2. Previous Work

Single-View DFD DFD is a well established field for single-view reconstruction. While some works achieve depth estimation from a single defocused image [43, 33], this is formally ill-posed. Instead, most existing literature use two or more defocused images taken from a single viewpoint and compare the relative blur between them [14]; simplifying the problem by not considering radiance when recovering shape [3, 22].

One approach for capturing these images is to vary the aperture setting [29, 25, 36]. Due to ambiguities in depth estimation, this constrains the scene to either in front of or behind the focused plane [24] - limiting the reconstruction volume. Alternatively, the camera can be refocused through the scene. While this approach has fewer constraints, it has the disadvantage of introducing scale and translational differences between images when captured with a conventional lens. Previous works have corrected for this computationally [38, 2] or optically [41]. However, the former approaches have reliance on scene content and DoF for accurate registration, while the latter requires modifying the camera. Some works instead use lightfield cameras to refocus the scene [39, 9, 40] at the expense of spatial resolution.

Regularisation is often introduced in textureless regions by formulating the problem as an energy function. This is typically implemented numerically [14, 3, 11, 22], or in an MRF framework [26, 30]. While numerical approaches have higher resolution reconstructions, they tend to lack the stability of MRF-based approaches. More recent works have started to explore the use of deep learning [36, 9, 7, 17], although this is still an emerging area of research.

Multi-View DFD While the majority of literature focuses on single-view reconstruction, some works have investigated the use of defocus cues with multiple viewpoints. These works combine DFD with correspondence information to improve stereo matching. As such, most do not use more than 2 views. [31] formulate DFD as disparity and apply stereo constraints. [21] estimate camera parameters and combine cues in a iterative framework. [8] formulate relative blur as a function of disparity and estimate the blurring function. [4] derive an expression to calculate relative blur between arbitrary viewpoints. [39] utilise a lightfield camera and combine correspondence and defocus information in an MRF framework.

Summary Very few approaches consider camera models other than the thin lens model. This includes most deep learning-based approaches, which typically adopt a thin lens when generating training data. Although [16] propose a data-driven camera model, they note an implicit dependence on the thin lens model. Methods which consider additional viewpoints either iteratively approximate camera parameters, or rely on insufficent calibration. Moreover, most do not consider more than 2 views. In contrast, we propose the use of a generalised camera model and calibration procedure that enables complete scene modelling from an arbitary number of multi-view focal stacks. Our approach is intended to be as generalised as possible, and as such does not require any modification of the camera or specialist equipment.

3. Camera Model

The basis of our model relies on the combination of projective geometry and finite-aperture camera models. In this section, we explain how this is applied to the formation of a focal stack from an arbitary viewpoint of the scene. From here onwards, we refer to parameters related to the i^{th} focus setting of this focal stack with a subscript. Without loss of generality, let us define a reference setting at $i = 0$.

For an ideal pinhole camera, the projection of a world-space coordinate $\boldsymbol{X}$ to an image-space coordinate $\boldsymbol{x}_i$ can be defined by the perspective transform [15]:

$$\boldsymbol{x}_i = K_i E \boldsymbol{X}, \tag{1}$$

where K_i and E are the camera intrinsic and extrinsic matrices respectively. For simplicity we assume images are not subject to lens distortion, but in reality we account for this during calibration. We define K_i as

$$K_i = \begin{bmatrix} s_i F_i & 0 & x_0 + \bar{t}_{ix} \\ 0 & s_i F_i & y_0 + \bar{t}_{iy} \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where x_0 and y_0 are the principal point, and F_i is the effective focal length. s_i and $\bar{\boldsymbol{t}}_{\boldsymbol{i}}$ account for the scale and translation differences relative to the reference setting, introduced by refocusing the camera. Let the radiance of $\boldsymbol{x}_i$ be defined by the function $r(\boldsymbol{x}_i)$. Defocus introduced to image I_i in the focal stack from the finite aperture is modelled by integrating the scene radiance over the shift-variant point spread function (PSF) of the camera k [14, 10, 13]:

$$I_i(\boldsymbol{y}) = \int k(\boldsymbol{y}, \boldsymbol{x}_i)\, r(\boldsymbol{x}_i)\, d\boldsymbol{x}, \tag{3}$$

where $I_i(\boldsymbol{y})$ defines the colour of pixel $\boldsymbol{y}$. Such a formation model only considers the behaviour of light acting as a particle, ignoring diffraction blur and chromatic aberration. However, given a large enough aperture and sufficently high quality lens, such effects can be considered negligble.

As in many works, we simplify Equation 3 to a convolution by assuming the scene is composed of fronto-parallel surfaces. Although this assumption becomes invalid at discontinuities [34] and more accurate models exist [35, 1, 12], we found the error introduced is insignificant for small blur radii. Thus, Equation 3 becomes [14, 10]

$$I_i(\boldsymbol{y}) = \int k_\sigma(\boldsymbol{y} - \boldsymbol{x}_i)\, r(\boldsymbol{x}_i)\, d\boldsymbol{x} \;=\; (k_\sigma * r)(\boldsymbol{y}). \tag{4}$$

In many previous works, k_σ is frequently approximated by either a Pillbox function [41, 11], resembling a circular PSF with hard edges; or a Gaussian function [14, 2], which approximates a multi-wavelength PSF [29, 4] as well as some diffraction effects [24, 11, 20]. In this work, we define k_σ as the latter

$$k_\sigma(\boldsymbol{y}) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{\boldsymbol{y}}{\sigma}\right)^2}. \tag{5}$$

To complete our camera model, we now need to define the blur radius σ. This is where our approach diverges significantly from other works. Typically, a thin lens model defines σ as a function of depth d [10]

$$\sigma(d) = \frac{\gamma a_i v_i}{2} \left(\frac{1}{d} + \frac{1}{v_i} - \frac{1}{f_i} \right), \tag{6}$$

where f_i is focal length, a_i is the aperture radius, v_i is the image distance and γ is a constant. Note that, unlike most works, we do not consider focal length and aperture size to be constant across all settings.

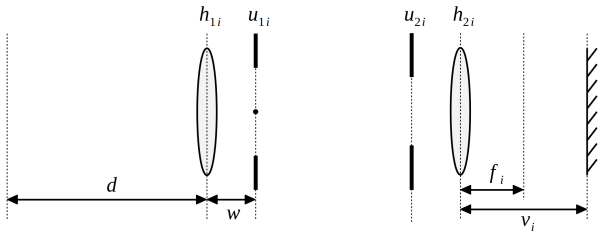


Figure 2. Our camera model is a thick lens composed of two thin lenses with focal length f_i separated by some distance. The effective pinhole location is at the entrance pupil u_{1i}. Calculation of the defocus radius σ for a given pixel is performed relative to the principal planes h_{1i} and h_{2i}.

By definition, Equation 6 implicitly assumes the principal plane (i.e. the thin lens) where light is modelled as refracting is aligned with the effective camera pinhole. We have observed this is not normally the case in reality, especially with macro lenses. This highlights a significant limitation with the standard thin lens assumption, particularly when using multiple viewpoints.

Instead, we adopt a thick lens model as seen in Figure 2 which accounts for this displacement. Our model is composed of two principal planes h_{1i} and h_{2i}. The amount of light entering or leaving each lens is controlled by the diameters of the entrance u_{1i} and exit u_{2i} pupils respectively. These are virtual images of the physical aperture as viewed from the front and the back of the lens. The effective pinhole is located at u_{1i}. Given the pupil ratio p_i [32]

$$p_i = \frac{u_{2i}}{u_{1i}}, \tag{7}$$

we define the displacement of the reference setting w [32]

$$w = f_0 \left(\frac{1}{p_0} - 1 \right). \tag{8}$$

Note as $p \to 1$, our model converges to a thin lens. Equation 6 is modified to become

$$\sigma(d) = \frac{\gamma a_i v_i}{2} \left(\frac{1}{d - w} + \frac{1}{v_i} - \frac{1}{f_i} \right). \tag{9}$$

4. Calibration

The calibration of our camera model introduced in Section 3 is non-trivial for several reasons. First, unlike most approaches, we do not consider relevant camera parameters provided by the manufacturer to be accurate for all focus settings. This is because these values are only valid when the camera is focused at infinity. Secondly, to our knowledge there is no standard approach for reliably calculating the pupil ratio, whose value is of significant importance in our model. Finally, our calibration needs to correct for

Figure 3. Calibration images of a uniform plane used for deriving average brightness focused at infinity (left) and at a focus setting (right). The observed change in brightness is purely a result of refocusing the camera. Images are white balanced and brightened for visualisation.

translation and scale differences between multi-focus images without dependance on DoF or texture content.

In this section, we will discuss how we solved these problems. We begin by defining a number of focus settings that sweep through the scene volume. In general, the more focus settings captured, the better our model can be applied to DFD reconstruction. Our calibration approach can then be broken down into several stages. For each setting, the following key steps are made:

1. Calculate camera intrinsics and lens distortion
2. Derive affine transforms to register images
3. Estimate the defocus parameters in our model
4. Refine parameters in a per-viewpoint optimisation

4.1. Camera Matrices

In this first step, we derive the intrinsic calibration of the camera using a standard approach proposed in [42]. A calibration plane is positioned in multiple orientations and captured at each focus setting. Images are taken with both a small and a large aperture. For each setting, feature points are identified from the smaller aperture images. The intrinsic matrix K_i and lens distortion coefficents are solved by minimising the reprojection error. In the following sections, images have lens distortion removed. E is calculated in a similar way for each viewpoint, using a set of scene features common to all views.

4.2. Registration

This step aims to register all images in a focal stack to a reference setting. A naive approach may be to directly use the parameters from the geometric calibration. Since F_i is related to the projection magnification m_i by [32]

$$F_i = f_i \left(1 + \frac{m_i}{p_i}\right), \tag{10}$$

the scaling between two settings could be found quite easily if $p_i = 1$ and $f_i = f \; \forall \; i$. However, in our model neither of these conditions are guaranteed. In addition, while translation differences could be derived from the principal point in theory, in practise the estimation of this quantity is ill-posed and subject to unpredictable variations.

Instead, we exploit the detected features $\boldsymbol{c}$ from Section 4.1. By identifying corresponding features in the images, an optimal scale and translation can be calculated to best align them. The ratio of effective focal lengths between the reference F_0 and F_i is used as an initial scaling factor s_i. This is refined in a least mean square optimisation:

$$\min_{s_i} \sum_k || \, \boldsymbol{t}_i^k - \bar{\boldsymbol{t}}_i \, ||^2 \tag{11}$$

$$\boldsymbol{t}_i^k = \boldsymbol{c}_0^k - s_i \boldsymbol{c}_i^k \tag{12}$$

where $\boldsymbol{c}_0$ and $\boldsymbol{c}_i$ are the feature coordinates, and $\bar{\boldsymbol{t}}_i$ is the mean of $\boldsymbol{t}_i^k \; \forall \; k$. Equation 11 is solved as a function of s_i using gradient descent. Once s_i has been optimised, the corresponding $\bar{\boldsymbol{t}}_i$ represents the required 2D translation. After registration, all images in the focal stack share the camera matrices of the reference setting.

4.3. Parameter Estimation

In this section, we discuss how the parameters in Equation 9 f_i, a_i, v_i and w are estimated. We begin by calculating two intermediate variables m_i and p_i.

Pupil Ratio We capture images of a uniform plane focused at infinity (where $p_\infty = 1$), and at each of the defined focus settings (where p_i is unknown) as seen in Figure 3. From this, we find an expression relating the observed change in brightness as a result of refocusing to p_i. Here, we show the derivation of p_i when $u_2 < u_1$.

The amount of light incident to the image plane of the camera is related to the area of the smallest pupil (in this case u_2), with the global illumination and exposure time being the constants of proportionality. If they remain fixed, then the following must hold true:

$$\frac{b_\infty}{b_i} = \left(\frac{u_{1\infty}}{u_{1i} \, p_i}\right)^2. \tag{13}$$

Here, b_∞ and $u_{1\infty}$ are the average brightness and entrance pupil diameter focused at infinity; and b_i and u_{1i} are the average brightness and entrance pupil diameter at a given focus setting. Note that $u_{1\infty} = u_{2\infty}$ and $u_{1i} \; p_i = u_{2i}$. Equation 13 can be rewritten as:

$$\frac{b_\infty}{b_i} = \left(\frac{f_\infty N_i}{F_i N_\infty p_i}\right)^2. \tag{14}$$

Here, f_∞ is the known focal length when focused at infinity, N_∞ is the reported f-stop of the aperture and N_i is the

effective f-stop setting. Given that [32]

$$N_i = N_\infty \left(1 + \frac{m_i}{p_i}\right), \tag{15}$$

Equation 14 can be rearranged as a quadratic function of p_i by substituting Equation 15:

$$\frac{F_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}p_i^2 - p_i - m_i = 0. \tag{16}$$

The value of p_i when $u_2 < u_1$ is given by the positive solution of Equation 16. Note here that $p_i < 1$:

$$p_i = \frac{f_\infty}{2F_i\sqrt{\frac{b_\infty}{b_i}}}\left(1 + \sqrt{1 + \frac{4F_i m_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}}\right). \tag{17}$$

A similar derivation can be made for $u_2 > u_1$. Conversely, in this case $p_i > 1$:

$$p_i = \frac{m_i}{\frac{F_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}} - 1}. \tag{18}$$

The choice of either Equation 17 or 18 when calculating p_i is simply a case of whichever one gives a valid solution. The only unknown here is m_i, which we derive next.

Magnification The magnification m_i of a focus setting is found by first finding the focusing distance d_i. This is the distance from the camera pinhole to the centre of the DoF. m_i and d_i are related as follows [18]

$$m_i = \frac{F_i}{d_i}. \tag{19}$$

To calculate d_i, we apply the Sum Modified Laplacian (SML) [27] focus measure to the large aperture calibration pattern images captured in Section 4.1. Since the poses of the patterns are known, feature points on the calibration plane can be sampled and the distance to the camera found. Regions where a high response is measured indicates an area in-focus. Assuming the DoF is a parallel plane, samples from multiple calibration images can be collected to improve robustness. The weighted mean of the distribution above a threshold gives the value of d_i as illustrated in Figure 4, from which m_i is found.

Focal Length Given m_i, p_i and F_i, the value of f_i is given by rearranging Equation 10 as

$$f_i = \frac{F_i}{\left(1 + \frac{m_i}{p_i}\right)}. \tag{20}$$

Aperture The aperture radius a_i is given by [18]

$$a_i = \frac{F_i}{2N_i}. \tag{21}$$

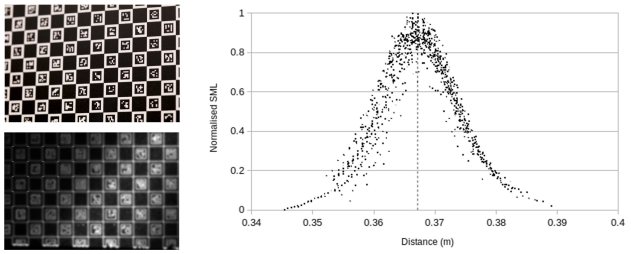


Figure 4. Left: Finite-aperture calibration plane image with known pose (top) and SML focus response with perspective distortion removed (bottom). Sample points are indicated by white circles. Right: distribution of focus samples accumulated from all calibration images used to calculate magnification. Weighted mean indicated by dashed line.

Image Distance Usually, v_i is defined by [18]

$$v_i = f_i(1 + m_i). \tag{22}$$

While this is correct for a single image, this does not hold in the context of a focal stack. This is because, as the camera is refocused, the principal planes do not remain in a fixed position. Thus, for DFD observations to be relative to the same point (the reference focus setting at $i = 0$), this drift needs to be accounted for when calculating v_i

$$v_i \;=\; f_i(1 + m_i) - (f_0 - f_i) \;=\; f_i(2 + m_i) - f_0. \tag{23}$$

Equation 23 offsets Equation 22 by the difference in focal length relative to f_0. This is illustrated in Figure 5 using a thin lens for simplicity. Essentially, this adjustment ensures the principal planes of each setting align with one another.

Pinhole Offset Finally, we can now define the value of w according to Equation 8.

4.4. Parameter Refinement

An important practical consideration during acquisition is to capture multiple focal stacks with the same settings. So far, we have assumed the ideal case where the camera refocuses perfectly. However, throughout the calibration process the lens will not be returning to exactly the same focus setting. As a result, there may be a need to refine some parameters on a per-viewpoint basis, depending on the quality of the lens. In our experience, only the value of w needs adjusting in this way. All other parameters (including registration) are deemed sufficently accurate.

We optimise w using scene features with known position in the world reference frame. Our cost function is based on the relative blur between pairs of images in the focal stack. The cost function presented here is similar to the one used in Section 5 for reconstruction. First, we define the relative blur between settings i and j:

$$\sigma_{ij}(d) = \sqrt{|\sigma_i(d)^2 - \sigma_j(d)^2|} \tag{24}$$

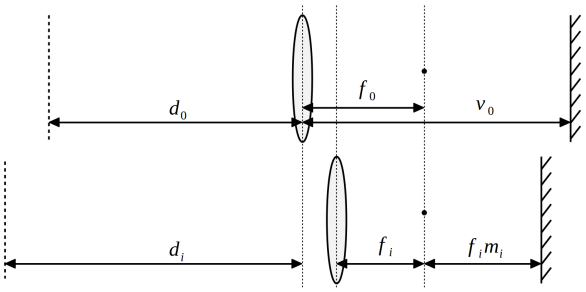


Figure 5. Thin lens illustration demonstrating how refocusing through the scene offsets the principal plane. Relative to the reference setting (top), the effective position of the lens for some other setting focused further away (bottom) is displaced by the difference in focal lengths. Note this is not related to w.

where $\sigma(d)$ is defined in Equation 9. Using this, we optimise w using images I_a and I_b from the focal stack.

$$\min_{w} \sum_{\{ij\}\in\mathbf{\Omega}} \sum_{k} ||\sigma_{ij}(d^k) \circ I_a \; - \; I_b \;||^2 \tag{25}$$

$$\{a,b\} \; = \; \begin{cases} \{i,j\} & \sigma_i(d) < \sigma_j(d) \\ \{j,i\} & \text{otherwise} \end{cases} \tag{26}$$

Here, $\mathbf{\Omega}$ is a vector of paired image indices, $\circ$ is a defocus operator which we define later, and d^k is the distance of the k^{th} feature from the camera. Equation 25 blurs whichever image is sharpest to match the other for every feature, and compares the result with a pixel-wise square difference. This sparse optimisation can be thought of as a per-viewpoint global adjustment of all blurring functions describing the focal stack.

5. Reconstruction

Our reconstruction framework is defined as a view-dependent, MRF-based discrete labelling problem. Each pixel in a given image represents the appearance of a scene surface, which we model as a tangent plane. Thus, to reconstruct the scene, we solve the inverse problem of finding the world-space position and orientation of every surface using the defocus information leveraged from our camera model.

We propose an iterative framework which overcomes the limited depth resolution inherent to MRF-based implementations. For a given viewpoint, let us define a volume *unique to each pixel* within which the corresponding surface is located. These volume boundaries could be given, for example, by a visual hull initialisation. Then, we uniformly divide these volumes into N candidate labels $\boldsymbol{x} = \{x_1, x_2, \; ... \; x_N\}$; where each label corresponds to a unique distance from the camera. For every pixel, we derive costs Φ_D associated with the likelihood of label assignment from the observed defocus across the focal stack. The optimal labelling is then found from the energy function

$$E(\boldsymbol{x}, n) = \sum_{\boldsymbol{p}\in\nu} \Phi_D(x_{\boldsymbol{p}}) \; + \; \frac{\lambda}{n} \sum_{\{\boldsymbol{p},\boldsymbol{q}\}\in\epsilon} \Psi_{\boldsymbol{pq}}(x_{\boldsymbol{p}}, x_{\boldsymbol{q}}). \tag{27}$$

In untextured or ambiguous regions, $\Psi_{\boldsymbol{pq}}$ encourages a labelling consistent with neighbouring surfaces according to the value of λ; whose effect reduces with iteration n to encourage higher fidelity. We define one iteration in this context as a complete optimisation of Equation 27 for all labels using α-expansion [37, 6, 19, 5].

After each iteration, our novel approach reduces the reconstruction volume asscociated with every pixel by half around the current labelling. Since the number of candidate labels N remains the same, the resolution of the framework doubles with every iteration. This can be considered a globally optimal prunning of unnecessary labels, and signifcantly reduces memory usage and computational cost while simultaneously offering limitless depth resolution. Surface normals are updated from the labelling gradient and used by $\Psi_{\boldsymbol{pq}}$ in the next iteration.

Once the target resolution is reached, a point cloud is generated from the depth and normal maps, and combined with point clouds from other views. Minimal filtering and downsampling of the point cloud is performed, before a surface mesh is generated using standard Poisson surface reconstruction. The remainder of this section will discuss the terms in Equation 27 in more detail.

5.1. Defocus Term

To calculate the defocus term for a pair of images $\{I_i, I_j\}$ in a given focal stack, a scale-space approach is taken. The relative blur between the images is found according to Equation 24, and the sharper image is blurred to match the other. The cost function $\phi_D(x_{\boldsymbol{p}})$ is defined by the square difference between the defocused and original image

$$\phi_D(x_{\boldsymbol{p}}) = \sum_{\{ij\}\in\mathbf{\Omega}} \sum_{k} \left(\sigma_{ij}(d^k) \circ I_a \; - \; I_b \;\right)^2. \tag{28}$$

As in Equation 25, $\circ$ denotes the defocus operator, $\mathbf{\Omega}$ contains indices of paired images, and $\{a,b\}$ are defined in Equation 26. Since the accuracy of DFD is greatest when relative blur is small, only neighbouring images in the stack are paired together. When evaluating Equation 28, we first remove harmonic texture components in the source images

$$I_i = I_i - (I_i \circ k_\sigma). \tag{29}$$

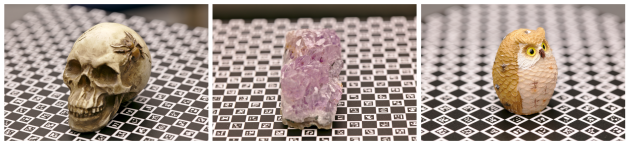


Figure 6. Example images from the presented datasets.

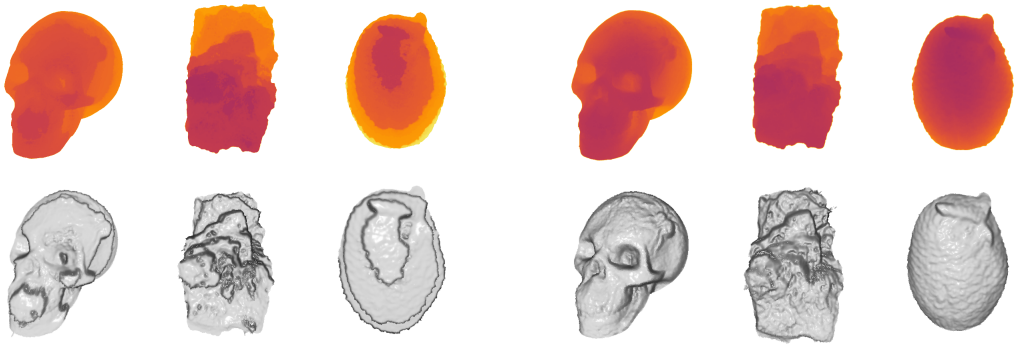


Figure 7. Example single view reconstructions of our datasets. The depth maps (top row) and corresponding point clouds (bottom row) are recovered using the thin lens model (left) and our thick lens model (right).

This procedure, proposed in [10], removes defocus-invariant texture components, and has been shown to improve the performance of DFD. We define our defocus operator $\circ$ as a linear diffusion operator as proposed in [14]

$$I_i \circ \sigma = I_i + c_\sigma \left(\nabla^2 * I_i\right), \tag{30}$$

where ∇^2 denotes the Laplacian operator, and c_σ is termed the diffusion coefficent

$$c_\sigma = \frac{\sigma^2}{2t}. \tag{31}$$

Here, we set diffusion time $t = 0.5$ as suggested by [14]. Although $\circ$ is equivalent to a convolution with k_σ, we found linear diffusion performs better with subpixel defocus radii. The forward diffusion constraint is enforced by starting Equation 28 at the label closest to the depth d_0 where the relative blur $\sigma_{ij}(d_0) = 0$. We derive this from Equation 24:

$$d_0 = \frac{a_i v_i \pm a_j v_j}{\frac{a_i}{f_i}(v_i - f_i) \pm \frac{a_j}{f_j}(v_j - f_j)} + w \tag{32}$$

The above simplifies to the result in [14] when $f_i = f_j$, $a_i = a_j$ and $w = 0$. Finally, the generated costs are normalised by the following, where μ_D is the mean of the unnormalised cost volume.

$$\Phi_D(x_{\boldsymbol{p}}) = 1 - e^{-\frac{\phi_D(x_{\boldsymbol{p}})}{\mu_D}} \tag{33}$$

5.2. Smoothness Term

Since the defocus term has a reliance on scene texture for accurate depth estimates, regularisation is necessary to ensure consistently smooth reconstructions in textureless regions. We define this by the function V, which we truncate to preserve discontinuities:

$$\Psi_{\boldsymbol{pq}}(x_{\boldsymbol{p}}, x_{\boldsymbol{q}}) = \min\left(\Psi_{max}, V_{\boldsymbol{pq}}(x_p, x_q)\right) \tag{34}$$

In our implementation, smoothness costs are calculated according to pairwise interactions over a 4-connected clique. As proposed by [28], we implement V as a second-order smoothness prior, by exploiting the surface model previously described. This encourages a piecewise linear reconstruction, rather than a fronto-parallel one. In combiniation with our data term and iterative label pruning, this makes for a very powerful framework for general scenes.

6. Evaluation

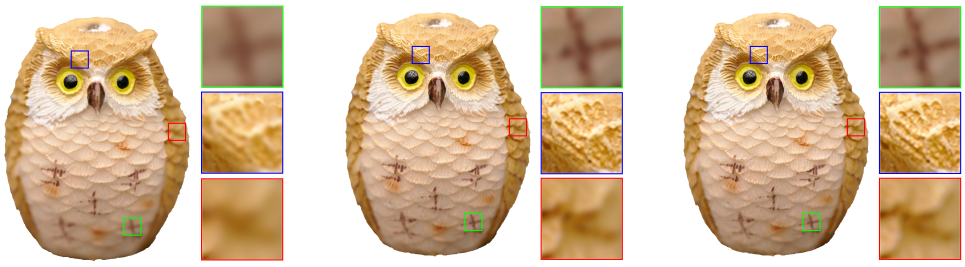
We now present an evaluation of our implementation, and compare the thick lens model to the traditional thin lens model. For the sake of fairness, both models are evaluated using the same DFD framework introduced in Section 5. We are unable to directly compare to other DFD formulations, since source code is not available from recent works that would produce meaningful comparisons. Additionally, public DFD datasets lack our thick lens calibration.

We therefore present 3 real-world multi-view datasets in this paper: Owl (29 views, 5 settings), Skull (16 views, 7 settings) and Quartz (7 views, 7 settings); see Figure 6. Each dataset was taken with a Canon EOS 5D camera using a 100mm macro lens. Ground truth geometry is unavailable as a result of their scale and surface complexity.

Table 1 shows some example parameters derived for each model. We precisely calibrate the thin lens using the intrinsic calibration from Section 4.1 to find v_i and d_i, as defined by the model. However, applying the thin lens assumptions to Equations 20 and 21 gives erroneous results, so we follow previous works in setting f_i and a_i to the values reported by the camera. Unlike our model, the thin lens reconstructions cannot utilise a visual hull initialisation.

Model	f (mm)	a (mm)	v (mm)	w (mm)
Thin	100.00	8.93	168.23	0.00
Thick	98.13	8.76	143.43	53.90

Table 1. Parameters for the reference setting of the Owl dataset.



Thin Lens Thick Lens Ground Truth

Figure 8. Reblurring results on one view in the Owl dataset. Using the reconstruction result from the thin and thick lens models, we have synthetically defocused an estimate of the scene radiance and compared to the corresponding image from the captured focal stack.

Dataset	Model	Focus Setting PSNR (dB)						
		0	1	2	3	4	5	6
Owl	Thin	30.80	32.62	32.31	29.32	28.75	-	-
	Thick	**35.83**	**39.28**	**38.90**	**35.07**	**32.65**	-	-
Skull	Thin	31.78	32.81	32.27	32.59	31.49	30.42	29.64
	Thick	**35.51**	**37.10**	**37.88**	**37.57**	**35.69**	**34.31**	**32.81**

Table 2. Results of the comparison between the synthetically generated and the captured images. These results were generated from a single view and across all focus settings for each dataset. In all cases, the thick lens model outperforms the thin lens model.

Our framework produces view-dependent reconstructions as seen in Figure 7. We assess the accuracy of the camera models by synthesising reblurred images using the recovered depth, and comparing to the captured images.

The effectiveness of our thick lens model is shown quantitively in Table 2 and qualitatively in Figure 8. From our results, it is clear our thick lens model and calibration outperforms the thin lens model in all tested cases. While the poor performance of the thin lens model may be exaggerated in our framework due to its iterative nature, these results illustrate its inherent limitations. By generalising the camera to a thick lens under the same conditions, reconstruction accuracy, and therefore defocus modelling, significantly improves. In a multi-view context, the thin lens reconstructions in Figure 9 do not coincide with one another. In contrast, our model produces consistent point clouds from multiple views as seen in Figure 1; supporting our hypothesis. Additional results can be found in supplementary work.

7. Conclusion

In this paper, we have demonstrated the limitations imposed by the thin lens camera model. Despite its success in DFD literature, it has restricted many previous works to coarse, single-view reconstructions that are not geometrically consistent. Here, we propose an alternative approach by adopting a thick lens model. Using our novel calibration procedure, we accurately model defocus formation across a registered focal stack. We apply this model to generating high quality 3D reconstructions of complex materials; something that was not previously feasible. The results shown verify our calibration approach, and demonstrate a significant improvement over the traditional model. We foresee our model being applied to reconstructing even smaller scenes, where the performance of traditional defocus modelling degrades further still; and potentially translating to other fields such as microscopy.

In our model, we assume defocus formation is consistent with a Gaussian convolution, which is not the case in reality. In future work, it would be interesting to explore the benefits of accurate PSF modelling; either with coded aperture or by directly measuring the camera PSF. We also intend to incorpoate deep learning to further improve the performance of our camera model.

Acknowledgments This research was supported by the EPSRC (grant EP/N509772/1).

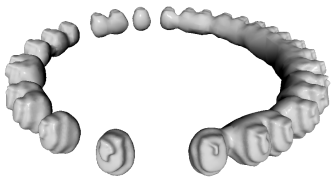


Figure 9. Multi-view reconstruction of the Owl dataset using the thin lens model. Due to the inaccurate defocus modelling, none of the individual point clouds intersect to form a coherent surface.

References

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

[2] R. Ben-Ari. A unified approach for registration and depth in depth from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1041–1055, 2014.

[3] R. Ben-Ari and G. Raveh. Variational depth from defocus in real-time. pages 522–529. IEEE Publishing, 2011.

[4] A. Bhavsar and A. Rajagopalan. Towards unrestrained depth inference with coherent occlusion filling. *International Journal of Computer Vision*, 97(2):167–190, 2012.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124,1137, 2004-09.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[7] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat. Deep depth from defocus: how can defocus blur improve 3D estimation using dense neural networks?, 2018.

[8] C.-H. Chen, H. Zhou, and T. Ahonen. Blur-aware disparity estimation from defocus stereo images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 2015, pages 855–863. IEEE, 2015.

[9] Z. Chen, X. Guo, S. Li, X. Cao, and J. Yu. A learning-based framework for hybrid depth-from-defocus and stereo matching, 2017.

[10] P. Favaro. Shape from focus and defocus: Convexity, quasiconvexity and defocus-invariant textures. pages 1–7. IEEE, 2007.

[11] P. Favaro. Recovering thin structures via nonlocal-means regularization with application to depth from defocus. pages 1133–1140. IEEE Publishing, 2010.

[12] P. Favaro and S. Soatto. Seeing beyond occlusions (and other marvels of a finite lens aperture). In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, volume 2, pages II–II. IEEE, 2003.

[13] P. Favaro and S. Soatto. A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):406–417, 2005.

[14] P. Favaro, S. Soatto, M. Burger, and S. Osher. Shape from defocus via diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):518–531, 2008.

[15] R. Hartley. *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK ; New York, 2000.

[16] S. Hasinoff and K. Kutulakos. Confocal stereo. *International Journal of Computer Vision*, 81(1):82–104, 2009.

[17] M. Kashiwagi, N. Mishima, T. Kozakaya, and S. Hiura. Deep depth from aberration map. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4069–4078, 2019.

[18] R. Kingslake. *Optics in photography*. SPIE Press monograph ; PM06. SPIE, Bellingham, Wash. (1000 20th St. Bellingham WA 98225-6705 USA).

[19] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147,159, 2004-02.

[20] I. Lee, M. Tariq Mahmood, and T.-S. Choi. Adaptive window selection for 3d shape recovery from image focus. *Optics and Laser Technology*, 45:21–31, 2013.

[21] F. Li, J. Sun, J. Wang, and J. Yu. Dual-focus stereo imaging. *Journal Of Electronic Imaging*, 19(4), 2010.

[22] X. Lin, J. Suo, X. Cao, and Q. Dai. Iterative feedback estimation of depth and radiance from defocused images. In *Computer Vision - ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV*, volume 7727 of *Lecture Notes in Computer Science*, pages 95–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[23] X. Lin, J. Suo, and Q. Dai. Extracting depth and radiance from a defocused video pair. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(4):557–569, 2015.

[24] F. Mannan and M. S. Langer. Optimal camera parameters for depth from defocus. In *2015 International Conference on 3D Vision*, pages 326–334. IEEE, 2015.

[25] M. Martinello, A. Wajs, S. Quan, H. Lee, C. Lim, T. Woo, W. Lee, S.-S. Kim, and D. Lee. Dual aperture photography: Image and depth from a mobile camera. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1,10. IEEE, 2015-04.

[26] V. P. Namboodiri, S. Chaudhuri, and S. Hadap. Regularized depth from defocus. pages 1520–1523. IEEE, 2008.

[27] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(8):824–831, Aug. 1994.

[28] C. Olsson, J. Uln, and Y. Boykov. In defense of 3d-label stereo. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1730–1737, 2013.

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

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

[31] A. N. Rajagopalan, S. Chaudhuri, and U. Mudenagudi. Depth estimation and image restoration using defocused stereo pairs. *IEEE transactions on pattern analysis and machine intelligence*, 26(11), 2004.

[32] A. Rowlands. Fundamental optical formulae. In *Physics of Digital Photography*, 2053-2563, pages 1–1 to 1–62. IOP Publishing, 2017.

[33] Z. H. Saeed Anwar and F. Porikli. Depth estimation and blur removal from a single out-of-focus image. In G. B. Tae-Kyun Kim, Stefanos Zafeiriou and K. Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 113.1–113.12. BMVA Press, September 2017.

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

[35] N. Shroff, A. Veeraraghavan, Y. Taguchi, O. Tuzel, A. Agrawal, and R. Chellappa. Variable focus video: Reconstructing depth and video for dynamic scenes. In *2012 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2012.

[36] G. Song and K. M. Lee. Depth estimation network for dual defocused images with different depth-of-field. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1563,1567. IEEE, 2018-10.

[37] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):1068–1080, 2008.

[38] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos. Depth from defocus in the wild. volume 2017-, pages 4773–4781. IEEE, 2017.

[39] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *2013 IEEE International Conference on Computer Vision*, pages 673–680. IEEE, 2013.

[40] M. W. Tao, P. P. Srinivasan, S. Hadap, J. Malik, and R. Ramamoorthi. Shape estimation from shading, defocus, and correspondence using light-field angular coherence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):546–560, 2017.

[41] M. Watanabe and S. Nayar. Rational filters for passive depth from defocus. *International Journal of Computer Vision*, 27(3):203–225, 1998.

[42] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[43] S. Zhuo and T. Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011.