Uniting Stereo and Depth-from-Defocus: A Thin Lens-based Variational Framework for Multiview Reconstruction

Robert D. Friedlander Huizong Yang Anthony J. Yezzi
Georgia Institute of Technology, School of Electrical and Computer Engineering
{rfriedlander3, huizong.yang}@gatech.edu, anthony.yezzi@ece.gatech.edu

Abstract

The problem of reconstructing three-dimensional (3D) scene geometry and radiometry from images is an important problem in computer vision and has applications in a variety of fields such as medicine and artifact preservation. However, state-of-the-art multiview algorithms assume a pinhole camera that incorrectly models defocus blur as a property of the scene instead of a property of the imaging process. We address the problem of dense 3D shape reconstruction from multiple viewpoints in situations where the image data exhibits noticeable defocus. We develop a mathematical framework for a fully generative variational algorithm that iteratively deforms an estimate of the foreground surface shapes and scene radiance such that irradiance estimates given by the thin lens forward model are photometrically consistent with the actual image data. This framework is founded on novel geometric computations of flux differentials across an evolving surface as well as gradients along occluding boundaries and their projections. While more work is needed to make them fit for practical use, the future potential of methods based on these computations is shown with experiments reconstructing simple object shapes from both synthetically generated and real defocused images. While our reconstruction algorithm has a higher computational cost than pinhole-based methods due to the more general optical model, it better reconstructs object proportions as well as sharp features that are blurred due to image defocus. As such, our geometry-based method provides a unified framework that extends the applicability of multiview reconstruction techniques to the poorly supported domain of defocused images where state-of-the-art pinhole-based methods fail.

1. Introduction

The problem of acquiring knowledge about the 3D geometry of a scene from images has been a central pillar in computer vision research for decades, and techniques for solving this problem have been applied to a variety of fields including medicine, security, and artifact preservation. The literature regarding this problem has largely been segmented into the one-cue "Shape-from-X" classes of methods, including multiview stereo and depth-from-defocus. While there have been many successful algorithms developed in both of these fields, they often suffer decreased performance when the assumption of one varying parameter is not met, which can easily happen outside of controlled laboratory conditions.

We propose here a novel variational framework for the dense reconstruction of scene shapes and radiances from a collection of defocused and calibrated images from multiple viewpoints. This framework provides the foundation for a class of unifying methods that allow for arbitrary combinations of different imaging cues. The vast majority of existing multiview reconstruction methods assume a pinhole which allows imaging to be modeled via perspective projection. However, pinholes cannot account for the limited depth of field of real lenses and can fail to accurately model the scene. In contrast, methods for depth-from-defocus use a more general optical kernel, but their fixed viewpoint does not allow them to model the scene in its entirety nor to exploit the information obtained from parallax. We propose a multiview stereo method that unifies these two approaches by utilizing a thin lens as the underlying camera model, allowing it to obtain information from both parallax and defocus. Our approach also includes novel solutions to difficult mathematical problems not well-addressed in the literature but that underscore the geometric nature of the reconstruction problem; these include computations of flux differentials across an evolving surface as well as gradients of occluding boundaries and their projections in the presence of defocus.

1.1. Related Work

Two primary approaches dominate the literature for multiview stereo: feature correspondence methods and variational methods. Correspondence methods utilize

epipolar geometry to find matching sets of image features and backtrace them to extract a 3D point cloud of the scene which must be further processed to form a connected shape [11, 32]. The quality of this point-matching is the main factor in the accuracy of the final reconstruction, and probabilistic methods have been employed in [4, 9, 10, 34] as well as iterative refinement in [33] to ensure optimal correspondences are obtained. However, correspondence methods work best for well-focused scenes due to their use of the pinhole model. When images exhibit noticeable defocus, the resulting homogeneous regions cause the correspondence problem to become ill-posed, resulting in a poor reconstruction.

In contrast, variational methods iteratively deform an initial surface estimate until it is photometrically consistent with the input images. The cost for the earliest variational methods for stereo by Faugeras and Keriven was the reprojection error, and later the cross-correlation, between different images at a collection of paired image points, wrapping correspondence methods in a variational framework [12, 13]. This was taken further by Yezzi and Soatto with the fully generative variational method developed in [51] where estimates of the scene radiance are used to generate synthetic images that can be directly compared to the corresponding image data, prompting greater numerical stability and robustness to specularities. Variational methods complement correspondence methods and work well for smooth objects and even high noise densities, but like correspondence methods they fail when the image data are out of focus. In addition to falsely treating defocus blur as a scene property due to the assumed pinhole model, these methods require (sometimes heavy) regularization on the surface area, causing them to be unable to reconstruct sharp corners and edges. In contrast, our method explicitly models defocus blur by assuming a thin lens model instead of a pinhole and allows for good reconstruction of non-smooth features without artificial regularization for sufficiently defocused imagery.

Depth-from-defocus steps out from the limitations of the pinhole model, either by using a thin lens model [1, 43] or a Gaussian or other estimated optical kernel [8, 38]. However, the limitation to a single viewpoint prevents depth-from-defocus methods from reconstructing the full scene, a limitation overcome by our multiview method. Our use of the thin lens model is inspired by these depth-from-defocus methods, and our work here can be seen as an extension of the variational methods in [14, 23] to the case of multiple viewpoints.

Some attempts have been made at applying deep learning to both multiview stereo [20, 42, 48, 50] and depth-from-defocus [5, 18], and a review of various deep learning-based reconstruction methods can be found in [19]. While effective, these methods still require ample training data and extensive training periods in order to perform well, whereas our proposed method only requires the images for the scene in question and can be used immediately.

There have been some previous attempts to integrate both stereo and defocus cues, but many of these methods either use point correspondence [47, 49] and thus share the same limitations as previous correspondence methods, or they require specialized equipment [39, 44, 45, 47] that prohibit general use. Also, most previous cue-fusion methods assume only one stereo image pair with short baseline, so only a single depth map is computed [3, 6, 31, 37, 39, 40, 44]. In contrast, our method works with off-the-shelf cameras and can process an arbitrary number of images and thus can fully recover object shape up to the field-of-view coverage given in the images.

The method proposed here is motivated by the generative variational methodology of Yezzi and Soatto in [51] and its many sequels [2, 7, 21, 22, 24, 25, 28, 26, 27, 29, 41, 46, 52, 53], and we attempt to relax their assumption of focused imagery by assuming a thin lens to account for image defocus, whose forward model is derived by Friedlander and Yezzi in [17]. It is akin to a generalized (with respect to imaging model) version of space carving [30] under a variational framework, with a key difference being that in our method voxels can be both added to and removed from the surface estimate whereas in space carving voxels can only be removed.

Finally, the implementation used to obtain the results herein utilized the level set methods of Osher and Sethian, which are well-known to be effective in numerically implementing evolving interfaces [35, 36]. It is also worth noting that flux differentials are briefly discussed in [15], but only for the specific case of rigid body motion, not deforming surfaces.

2. Variational Formulation

Following the philosophy from [51] and its sequels, we model the scene as a set of surfaces in space that each support a Lambertian radiance function. What differentiates our method is that instead of modeling the camera as the traditional pinhole, we model the camera using a thin lens, allowing us to account for defocus blur in images. We then apply the thin lens forward model to estimates of the surface shapes and radiances to produce corresponding image estimates to be compared to the input images. It is desired for the image estimates to be as close to the input images as possible, so we construct a cost functional that is minimized when the estimated surface shapes and radiances produce photometrically consistent images. This minimization is done using a gradient descent procedure.

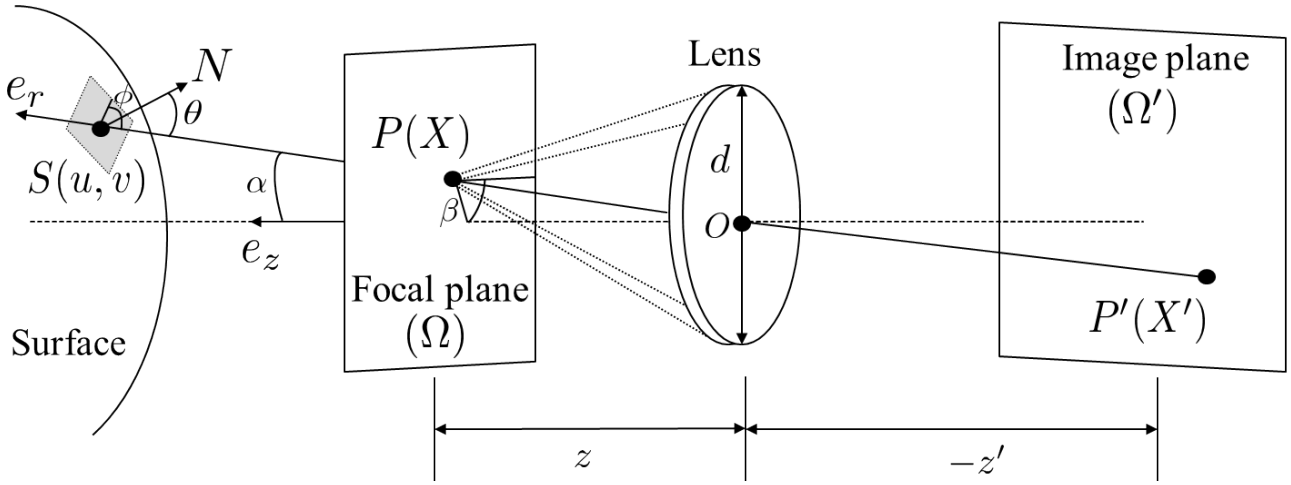

Figure 1. Visual depiction of thin lens imaging geometry and notation. The gray rectangle represents the tangent plane to the surface, and the dashed line represents the optical axis. While the ray shown here goes through the center of the lens O, there is a whole cone of rays that pass through both $P(X)$ and the lens as denoted by the dotted lines, and the size of this cone is determined by the focus depth z and lens diameter d. Adapted with permission from [16] © 2017 IEEE

2.1. Preliminaries and Notation

We denote by S a surface in $\mathbb{R}^3$ parameterized by coordinates (u, v) with area element dS, outward unit normal N, and supported radiance function $L : \mathbb{R}^2 \to \mathbb{R}$. The background of the scene is modeled as a second surface B that occupies the entire field of view under the "blue sky" assumption; we model B as a sphere parameterized by angular coordinates (η, γ) and supporting a radiance function $K : \mathbb{R}^2 \to \mathbb{R}$.

As previously mentioned, we model the camera as a circular thin lens whose center O is the origin of our coordinate system. The lens has focal length f and aperture diameter d. Given an image plane at (negative) depth z' behind the lens, all light emitted from a point on the focal plane at (positive) depth z in front of the lens and passing through the lens is perfectly focused to a point in the image plane. According to the thin lens model, the quantities f, z', and z are related according to

$$\frac{1}{f} = \frac{1}{z} - \frac{1}{z'} \tag{1}$$

Additionally, let $M = \frac{z'}{z}$ denote the lens magnification. Then a generic point on the focal plane $P(X) = (X, z)$ with planar coordinates $X = (x_1, x_2)$ and its corresponding point on the image plane $P'(X') = (X', z')$ with planar coordinates $X' = (x'_1, x'_2)$ are related by $X' = MX$. Similarly, the area elements of the focal and image planes, dX and dX' respectively, are related by $dX' = M^2 dX$. Image irradiance E', which is our estimate of image pixel intensity, is related to conjugate irradiance E in the focal plane by $E'(X) = M^{-2}E(X)$ [17]. Because of this one-to-one mapping of points and irradiances between the focal and image planes, we can model the imaging process as first forming a conjugate image in the focal plane (defined on the conjugate image domain Ω) and then mapping this image onto the image plane (defined on the corresponding image domain Ω').

A ray between a surface point $S(u, v)$ and focal plane point $P(X)$ is characterized by its length $r = \|S - P\|$ and direction $e_r = (S - P)/r$. These rays can be parameterized using either surface coordinates (u, v, θ, ϕ) or focal plane coordinates $(x_1, x_2, \alpha, \beta)$, where θ is the angle e_r makes with respect to N, ψ is the rotation angle of e_r about N, and α and β are the analogous ray angles with respect to the focal plane unit normal e_z. When integrating over sets of rays, it will be especially convenient to use solid angle elements $d\boldsymbol{\theta} = \sin\theta \, d\theta \, d\psi$ and $d\boldsymbol{\alpha} = \sin\alpha \, d\alpha \, d\beta$. Only light rays that pass through the lens contribute to the image, so we consider only this set of detected light rays, denoted by Γ. A visual depiction of this ray geometry is shown in Figure 1.

It is possible to rewrite integrals in surface coordinates as integrals in focal plane coordinates or mixed coordinates, and vice versa, using the following change of variables formulae:

$$d\boldsymbol{\alpha} = \frac{\cos\theta}{r^2} \, dS \tag{2a}$$

$$d\boldsymbol{\theta} = \frac{\cos\alpha}{r^2} \, dX \tag{2b}$$

Finally, given a focal plane point $P(X)$, we denote by $S^*_{(\Gamma_S(X))}$ (or more compactly S^*) the surface curve bounding the subpatch of S visible from $P(X)$. In general, the * superscript is used to denote the corresponding quantity at such a visibility boundary.

2.2. Thin Lens Forward Model

The conjugate irradiance E can be computed from the surface radiance L and background radiance K as

$$E(X) = \int_{\Gamma_S(X)} L(u,v)\cos\alpha\, d\boldsymbol{\alpha} + \int_{\Gamma_B(X)} K(\eta,\gamma)\cos\alpha\, d\boldsymbol{\alpha} \quad (3)$$

where $\Gamma(X) \subset \Gamma$ is the set of detected rays that pass through $P(X)$ in the focal plane, and $\Gamma_S(X) \subseteq \Gamma(X)$ and $\Gamma_B(X) \subseteq \Gamma(X)$ are the ray subsets emitted from only the surface and only the background respectively [17]. Here, the irradiance is computed by integrating the scene radiance over the entire cone of incident rays, rather than projecting the radiance down to the image along a single ray as done in pinhole-based stereo methods.

2.3. Cost function

Since we will be estimating the surface shape and the surface and background radiance functions using an iterative procedure, we can augment them to be time-varying so that they represent a class of evolving functions $S(u,v) \rightarrow S(u,v,t)$, $L(u,v) \rightarrow L(u,v,t)$, and $K(\eta,\gamma) \rightarrow K(\eta,\gamma,t)$, where t is an artificial gradient descent time parameter. Letting C be the total number of images and comparing the modeled image irradiance E'_c and the actual measured intensity I_c of the c-th image, we may construct the image error function

$$\mathcal{E}'_c(X'_c) = E'_c(X'_c) - I_c(X'_c) \quad (4)$$

where the subscript on X' is used to denote that these planar coordinates are with respect to the coordinate system of the c-th image. This same error function can be mapped to the conjugate domain Ω_c to produce an equivalent conjugate error function

$$\mathcal{E}_c(X_c) = M_c^{-2} E_c(X_c) - I_c(M_c X_c) \quad (5)$$

We then can set up the total squared-error cost function J for the full image collection as

$$J = \frac{1}{2}\sum_{c=1}^{C} \int_{\Omega'_c} (\mathcal{E}'_c)^2\, dX'_c = \frac{1}{2}\sum_{c=1}^{C} \int_{\Omega_c} M_c^2 \mathcal{E}_c^2\, dX_c \quad (6)$$

whose time derivative is

$$\frac{dJ}{dt} = \sum_{c=1}^{C} \int_{\Omega_c} \mathcal{E}_c \frac{dE_c}{dt}\, dX_c \quad (7)$$

Note that unlike in [51] our cost function (6) contains no artificial regularizers and only measures the image matching-error. If images are sufficiently defocused, there should be natural regularization that occurs through the use of the thin lens model. This is already hinted at in the form of (3), where the irradiance is a weighted average over all scene points that contribute to a respective image point, and will be even more evident in the evolution equation that is derived in Section 3.

We want to minimize our cost (6) with respect to S, L, and K, and this is done using an alternating gradient descent procedure. First, we fix an initial estimate for the surface shape S and find the optimal radiance functions L and K for this estimate. Then, we fix L and K and update S according to the gradient descent flow of J with respect to S. These two steps are then repeated until convergence. From (7) we can see that in order to find the sensitivity of J to perturbations in the surface shape and scene radiance, and thus obtain the desired gradient descent evolution equations, we need to find the corresponding sensitivities of the irradiance E.

3. Surface Evolution Equation

Here we develop the equation governing the desired evolution of S that minimizes our cost (6). In order to obtain this evolution, it is necessary to compute the flux differential across a deforming surface as well as specific gradients along the occluding surface boundary and its projection in the presence of defocus. The former gives the sensitivity of the image irradiance to perturbations of the surface while the latter allows for the order of integration in the cost derivative to be reversed. The derivation of these novel geometric equations is lengthy, so for the sake of space and clarity of notation only the final results of these computations are shown here and only with respect to a single image. However, it is straightforward to apply these results to the entire image collection.

3.1. Irradiance sensitivity

Using a change of variables (2a) and noting that $\cos\theta = -e_r \cdot N$, $\Gamma_S(X) \cup \Gamma_B(X) = \Gamma(X)$, and $\Gamma_S(X) \cap \Gamma_B(X) = \emptyset$, we may rewrite the first term of (3) as a flux integral:

$$E(X) = -\int_{S_{(\Gamma_S(X))}} \frac{\hat{Q}\cos\alpha}{r^2} e_r \cdot N\, dS + \int_{\Gamma(X)} K\cos\alpha\, d\boldsymbol{\alpha} \quad (8)$$

where $\hat{Q} \doteq \hat{L} - \hat{K}$, and $\hat{L} : \mathbb{R}^3 \rightarrow \mathbb{R}$ and $\hat{K} : \mathbb{R}^3 \rightarrow \mathbb{R}$ are volumetric extensions of L and K such that $\hat{L}(S) = L(S)$ and $\hat{K}(B) = K(B)$. If we assume that $\hat{L}$ and $\hat{K}$ are fixed and that only S evolves, then the time derivative of (8), after simplification, is

$$\frac{dE}{dt}(X) = \int_{S^*_{(\Gamma_S(X))}} \left(\frac{\hat{Q}^* \cos\alpha^* \kappa_r^*}{(r^*)^2 \sqrt{(\kappa_r^*)^2 + (\tau_r^*)^2}} \right) (S_t^* \cdot N^*) ds^* - \int_{S_{(\Gamma_S(X))}} \nabla \hat{Q} \cdot e_r \frac{\cos\alpha}{r^2} (S_t \cdot N)\, dS \tag{9}$$

where ds^* denotes the arclength element along the integration boundary S^*, and κ_r^* and τ_r^* are the normal curvature and geodesic torsion of S in the direction e_r^*, respectively. Note that the second term of (8) is independent of S and thus vanishes when taking the derivative. Also, only points on S^* that lie on an occluding boundary (*e.g.* at least one detected ray emitted from that point satisfies $e_r^* \cdot N^* = 0$) contribute to the contour integral in (9).

3.2. Total matching error sensitivity

Inserting the irradiance sensitivity (9) into the cost derivative (7) and swapping the order of integration allows us to extract the gradient descent flow for S as

$$S_t = \left[-\frac{\hat{Q}}{\sqrt{1 - (e_z \cdot N)^2}} \left(\int_{\Omega^*_{(\Gamma_S(S))}} \frac{\mathcal{E}^* \kappa_r^* \cos\alpha^*}{r^*}\, ds_X \right) + \nabla \hat{Q} \cdot \left(\int_{\Gamma_S(S)} \mathcal{E} \cos\theta\, e_r\, d\boldsymbol{\theta} \right) \right] N \tag{10}$$

where $\Omega^*_{(\Gamma_S(S))}$ is the line of points in the focal plane that make rays satisfying the occluding boundary condition with the surface point S, and ds_X is the arclength element in the focal plane.

Notice that each surface point is updated by a weighted averaging of the pointwise error $\mathcal{E}$ over the region of the image contributed to by that surface point. This is the key that leads to the natural regularization mentioned previously, as the impact of noise or other outlying measurements in the data will be averaged out, with the size of the averaging window correlating with the amount of defocus present in the images. Also, in the special case of modeling L and K as constant functions, the second term vanishes, meaning that only surface points on the occluding boundary need to be considered.

One final thing to note is that the first integral of (10) represents a diffusion term due to the presence of the radial curvature κ_r. Since both $\hat{Q}$ and $\mathcal{E}$ can have either positive or negative sign, this diffusion can be in the backwards direction and thus be numerically unstable. If the curvature term were outside the integral, we could simply set $\kappa_r = -1$, which would at least guarantee the update at each point is in the right direction since we know that κ_r is negative at occluding boundary points (since we are using an outward normal). But since κ_r is inside the integral, doing this normalization could potentially change the overall sign of the integral and thus the direction of the update. This can be avoided by keeping track of the sign of the actual gradient and flipping the sign of the normalized update if necessary, as the computation of the radial curvature is relatively inexpensive. However, in doing so we are no longer descending down the exact gradient of the cost function J, though we are traversing a shallower trajectory that will still lead to its minimum. Such a modification turns the diffusion into an advection, which can be stabilized with the proper choice of time step.

4. Scene Radiance Estimation

Once the surface has been updated according to (10), we need to update L and K as to be optimal with this new shape. In the general case of smooth radiance functions, this can be done by solving an optimization problem on the manifolds S and B, the specifics of which will be explored in future work. In the special case that the surface and background radiance functions are separately modeled as constant functions, which is the case considered in the experiments presented here, then the optimal values of L and K individually can be found by finding where the derivative of J vanishes, yielding

$$L_{\text{opt}} = \frac{\sum_{c=1}^{C} \left(\int_{\Omega_c} I_c W'_{S,c} - K W'_{B,c} W'_{S,c}\, dX'_c \right)}{\sum_{c=1}^{C} \int_{\Omega_c} W'^2_{S,c}\, dX'_c} \tag{11a}$$

$$K_{\text{opt}} = \frac{\sum_{c=1}^{C} \left(\int_{\Omega_c} I_c W'_{B,c} - L W'_{B,c} W'_{S,c}\, dX'_c \right)}{\sum_{c=1}^{C} \int_{\Omega_c} W'^2_{B,c}\, dX'_c} \tag{11b}$$

where $W'_{i,c} \doteq M^{-2} \int_{\Gamma_i(X_c)} \cos\alpha\, d\boldsymbol{\alpha}$ for $i \in \{S, B\}$. Note that (11a) and (11b) are coupled, so they need to be applied in an alternating fashion until steady state is reached. However, it was seen in practice that a good steady state approximation is usually reached after one application despite the coupling. It should be noted that this piecewise-constant radiance case assumes that the images themselves can be well-approximated as piecewise-constant, with well-defined foreground and background regions, reminiscent of Chan-Vese image segmentation but with the addition of a blurred transition region.

5. Experimental Results

There are few publically available datasets for combined stereo and depth-from-defocus, and these rarely contain the needed camera parameters for our algorithm, so custom datasets were acquired. Here we show the results of

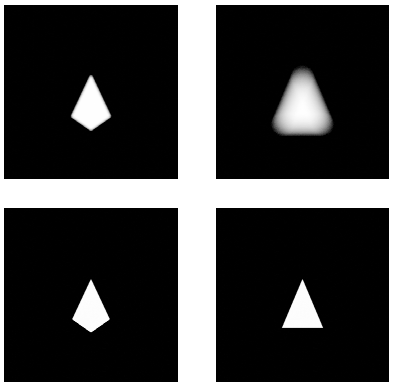

Figure 2. Lightly defocused (top left), heavily defocused (top right), and focused (bottom) images from the tetrahedron data set (2 of 30 views). Defocus blur makes sharp corners appear rounded and causes the tetrahedron to appear larger as the level of defocus increases

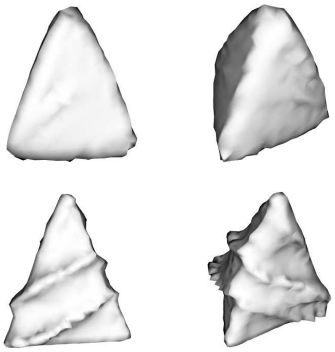

Figure 3. Two viewpoints of the thin lens-based reconstruction (top) and the pinhole-based reconstruction (bottom) for the tetrahedron data set. The thin lens method successfully reconstructed the single tetrahedron while the pinhole method could not reconcile the two different thicknesses visible in the images

testing our method on three different image sets, two photo-realistic sets obtained using Blender and one set obtained with a real camera. Due to space constraints, only a small selection of these images are displayed here. All experiments were run using an Intel i7-4790 processor with multi-threading, assumed that the radiance functions L and K were constant functions, and used a level set implementation with a $128 \times 128 \times 128$ voxel grid. The dimensions of the Blender-generated and real images were 480×480 and 576×384 respectively. For comparison purposes, we also applied the pinhole-based method of [51] to these data sets.

The first data set consisted of a tetrahedron imaged by 30 cameras placed in a 20 m diameter ring around and slightly above the tetrahedron. The first 15 cameras were lightly defocused with $f = 50\,\text{mm}$, $d = f/1.4$, and $z = 2\,\text{m}$, and the remaining 15 cameras were heavily defocused with $f = 50\,\text{mm}$, $d = f/1$, and $z = 500\,\text{mm}$. Such a situation could occur in an application combining stereo and depth-from-defocus. The top row of Figure 2 shows two of the images produced by these cameras while the bottom row shows focused versions of these views. As would be expected, the tetrahedron appears larger in all the defocused images due to defocus blur, and the edges and corners appear rounded instead of sharp; these effects are significantly more apparent in the heavily defocused images. The initial surface used was an ellipsoid containing the tetrahedron.

Figure 3 shows two viewpoints of the reconstructed models obtained using both the thin lens method and pinhole method. The pinhole method was unsuccessful and reconstructs the shape as a conglomeration of two tetrahedra of different thicknesses. This failure occurred because the pinhole model cannot reconcile the discrepancies between the two apparent sizes of the tetrahedron that can be seen in the images due to its implicit assumption of perfect focus. In contrast, our thin lens method treats defocus as an imaging property, so it was able to account for the different levels of defocus blur and reconstruct the tetrahedron as a single coherent shape.

To test our method's performance on rounded and partially concave shapes, the second data set consisted of a slanted dumbbell, with the cameras situated as before but this time all with identical focal parameters of $f = 50\,\text{mm}$, $d = f/1.4$, and $z = 500\,\text{mm}$. Such a situation could arise in a multiview application where the cameras are misfocused. The imaged scene is illustrated in Figure 4, with the left and right images corresponding to a focused image of the dumbbell and a defocused image actually used for the reconstruction. Again, the initial surface was an ellipsoid containing the dumbbell, and the resulting reconstructions from each method are shown in the left column of Figure 5. The thin lens reconstruction more accurately models the sharp edges of and the actual thickness of the disks and rod compared to the pinhole reconstruction. However, the pinhole reconstruction has a more appealing, though less accurate in terms of proportions, shape as it was able to fully carve out the area where the rod and disks intersect.

The experiment was repeated with noisy versions of the dumbbell images, and the resulting reconstructions can be seen in the right column of Figure 5. While the noise barely affected the quality of the thin lens reconstruction, it caused a noticeable decrease in the smoothness of the pinhole reconstruction even with a heavy smoothness penalty. This was due to the averaging that takes place in the surface update for the thin lens method but which is not present in the pinhole method.

To test our method's potential on real images, the third data set consisted of 32 images of a cube placed upon a thin

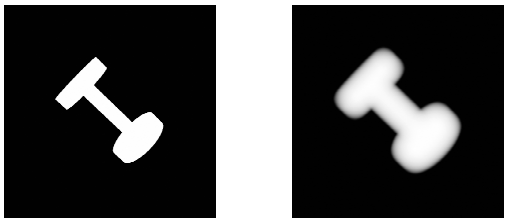

Figure 4. Focused (left) and defocused (right images from the dumbbell data set (1 of 30 views). Defocus blur makes the entire object appear wider, with less of a gap between the disks

trapezoidal-prism; a circle pattern was used for calibration, with the pattern being covered for the images used in the actual reconstruction. The camera (Nikon D3100 with included lens) was fixed on a tripod with the object on a turntable that was rotated to generate images at different viewpoints. One of these viewpoints can be seen in the top row Figure 7, with a focused view on the left and defocused view on the right. As with the Blender images, the sharp edges of the model appear rounded and elongated, and these blurred edges make the model's proportions appear larger. For the defocused images, the camera had $f = 32\,\text{mm}$, $d = f/5$, and $z = 173\,\text{mm}$. The initial surface was a box containing the model.

The obtained reconstructions in Figure 8 show that the thin lens method was able to reconstruct the model's shape and size more accurately than the pinhole method. The pinhole method resulted in a too-thick reconstruction where the cavity under the cube is only partially carved out; in contrast, the proper sizes of the model's shapes and the cavity can be seen in the thin lens reconstruction. The size difference is also seen in the silhouettes shown in the bottom of Figure 7. In addition, the thin lens reconstruction is smoother in several areas than the pinhole reconstruction, like the cube face shown in Figure 8, despite not having any explicit smoothness penalty. Four snapshots of the evolution process are shown in Figure 9.

Quantitative results for these experiments are summarized in Table 1. Due to the increased number of pixels needed to process for every updated surface point, from the line integral in the computation of S_t (10), the thin lens method is significantly more computationally expensive than the pinhole method. The above experiments

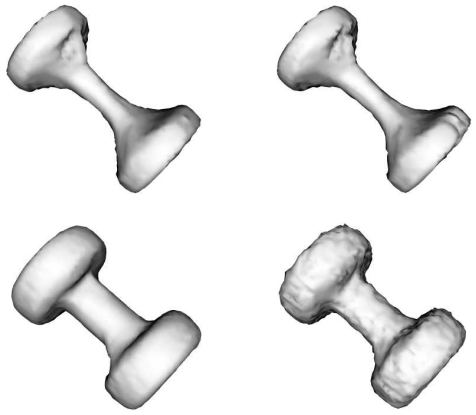

Figure 5. One viewpoint (left) of the thin lens-based reconstruction (top) and the pinhole-based reconstruction (bottom) for the dumbbell data set. The thin lens method more accurately reconstructs the sharp edges and actual thickness of the object compared to the pinhole method, but it gives a lower visual quality from its inability to fully carve out the rod-disk intersection. When noise was added (right), the thin lens reconstruction was barely affected while the pinhole reconstruction was noticeably more noisy and less smooth

had iteration times ranging from 1 to 3.5 minutes for the thin lens method while the pinhole method ran consistently at 100 ms or less per iteration. However, the thin lens method converged faster (in terms of the number of iterations) than the pinhole method for the tetrahedron and model data sets. Also, the relative increase in necessary iterations between the noiseless and noisy dumbbell images was significantly less for the thin lens method (80) than for the pinhole method (200). To measure the quantitative error of the final reconstructions relative to the input images, we used the total squared-error between the predicted (calculated using the thin lens forward model) and the actual pixel intensity. For comparison purposes, error for the noisy dumbbell reconstructions were computed with respect to the noiseless images. The thin lens method gave a noticeably lower error for all data sets.

6. Conclusion

We have developed a mathematical framework, based on novel geometric computations, for multiview reconstruction

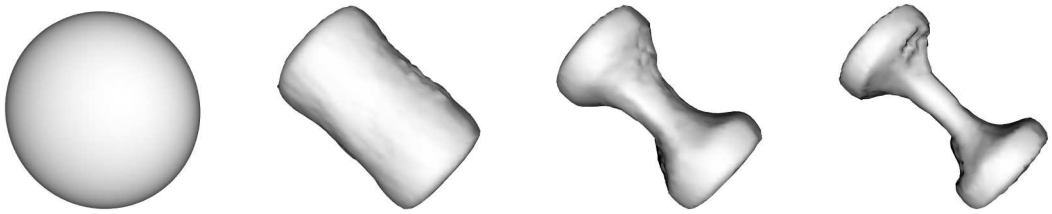

Figure 6. Four snapshots of the evolving surface for the dumbbell data set using the thin lens method after 0, 160, 240, and 400 iterations

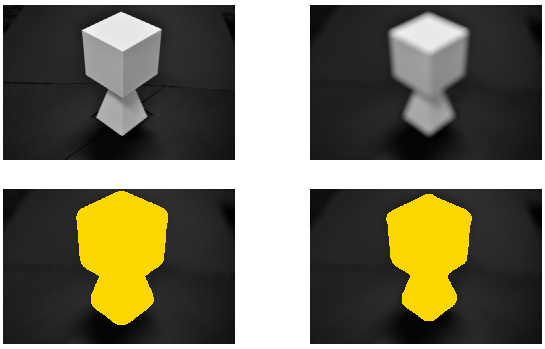

Figure 7. Focused (top left) and defocused (top right) images from the model data set (1 of 32 views). Sharp features like occluding edges are blurred outward, making the model appear larger and more rounded. The blurred silhouette of the pinhole reconstruction (bottom left) spreads out into the background while that of the thin lens reconstruction (bottom right) better matches the foreground region of the defocused images

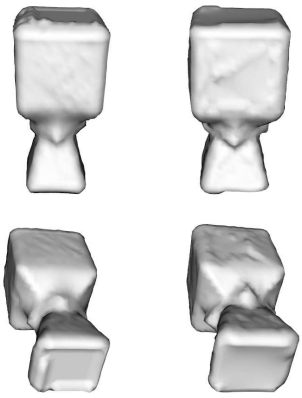

Figure 8. Two viewpoints of the thin lens-based (left) and pinhole-based reconstructions (right) for the model data set. Even in its current form, the thin lens method is able to more accurately reconstruct the size of the model than the pinhole method. Greater improvement is expected once more work has been done to fully develop this method for the more general smooth radiance case

of dense surfaces that is effective when the input images are defocused, a situation where current pinhole-based methods fail. It combines stereo with depth-from-defocus as the camera is modeled as a thin lens instead of a pinhole; this allows for the successful reconstruction of sharp edges and corners that appear rounded in images due to defocus blur, as a thin lens appropriately models defocus blur as a property of the imaging process and not of the scene. However, this performance comes at the cost of increased computation time with the need to integrate over a sizable number of pixels for every updated surface point each iteration. Even so, this very same integration grants our method a form of natural regularization that decreases the need for artificial regularizers, and this benefit is proportional to the level of defocus. Once fully developed, our thin lens-based method has the potential to be an effective complement to existing methods that can be applied to a large number of previously unsupported situations where the images are not well-focused or the one-cue assumptions of stereo and depth-from-defocus are not met.

Acknowledgements: This work was funded by the Army Research Office (W911NF-18-1-0281).

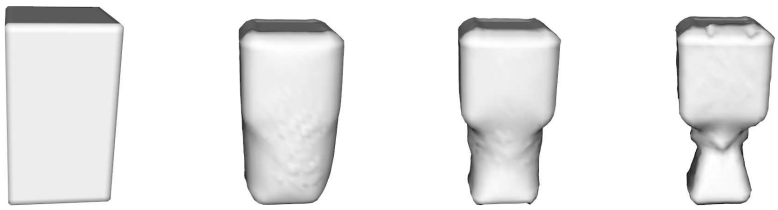

Figure 9. Four snapshots of the evolving surface for the model data set using the thin lens method after 0, 20, 40, and 74 iterations

Table 1. Performance Comparison of Thin Lens and Pinhole Reconstruction Methods

Data Set	Error		# Iterations		Max. Iter. Time (s)	
	Thin Lens	Pinhole	Thin Lens	Pinhole	Thin Lens	Pinhole
Tetrahedron	9.10×10^9	1.03×10^{10}	120	160	65	0.05
Dumbbell	1.91×10^9	7.11×10^9	400	200	217	0.1
Noisy Dumbbell	1.87×10^9	7.18×10^9	480	400	217	0.1
Model	4.34×10^9	1.17×10^{10}	74	170	51	0.1

The thin lens reconstructions were more photometrically consistent than the pinhole reconstructions, requiring fewer iterations to converge for half the datasets, including the real image one, even though each iteration itself was more costly

References

[1] Naoki Asada, Hisanaga Fujiwara, and Takashi Matsuyama. Edge and depth from focus. Int. J. of Comp. Vision, 26:153–163, 1998. 2

[2] Alvise Benetazzo, Francesco Fedele, Guillermo Gallego, Ping-Chang Shih, and Anthony Yezzi. Offshore stereo measurements of gravity waves. Coastal Engineering, 64:127–138, June 2012. 2

[3] Arnav V. Bhavsar and A. N. Rajagopalan. Towards unrestrained depth inference with coherent occlusion filling. Int J Comput Vis, pages 167–190, 2012. 2

[4] JiaWang Bian, Wen-Yan Lin, Yun Liu, Le Zhang, Sai-Kit Yeung, Ming-Ming Cheng, and Ian Reid. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. Int J Comput Vis, pages 1580–1593, Dec. 2020. 2

[5] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. Deep depth from defocus: How can defocus blur improve 3D estimation using dense neural networks? In Laura Leal-Taixé and Stefan Roth, editors, Computer Vision – ECCV 2018 Workshops, pages 307–323, Cham, 2019. Springer International Publishing. 2

[6] C. Chen, H. Zhou, and T. Ahonen. Blur-aware disparity estimation from defocus stereo images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 855–863, 2015. 2

[7] Navdeep Dahiya, Anthony Yezzi, Marina Piccinelli, and Ernest Garcia. Integrated 3D anatomical model for automatic myocardial segmentation in cardiac CT imagery. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pages 1–17, 03 2019. 2

[8] Trevor Darell and Kwangyoen Wohn. Depth from defocus using a pyramid architecture. Pattern Recognition Letters, 11:787–796, 1990. 2

[9] Frank Dellaert, Steven M. Seitz, Charles E. Thorpe, and Sebastian Thrun. Structure from motion without correspondence. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), volume 2, pages 557–564 vol.2, June 2000. 2

[10] Frank Dellaert, Steven M. Seitz, Sebastian Thrun, and Charles E. Thorpe. Feature correspondence: A Markov chain Monte Carlo approach. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems 13, pages 852–858. MIT Press, 2001. 2

[11] Olivier Faugeras. Three dimensional vision, a geometric viewpoint. MIT Press, 1993. 2

[12] Olivier Faugeras, Jose Gomes, and Renaud Keriven. Computational stereo: A variational method. In Geometric Level Set Methods in Imaging, Vision, and Graphics, pages 343–360. Springer-Verlag, New York, 2003. 2

[13] Olivier Faugeras and Renaud Keriven. Variational principles, surface evolution, PDEs, level set methods and the stereo problem. Technical Report 3021, INRIA, 1996. 2

[14] Paolo Favaro and Stefano Soatto. 3D Shape Estimation and Image Restoration: Exploiting Defocus and Motion Blur. Springer, 2007. 2

[15] Harley Flanders. Differentiation under the integral sign. *The American Mathematical Monthly*, 80(6):615–627, 1973. 2

[16] Robert D. Friedlander and Anthony J. Yezzi. A closed-form expression for thin lens image irradiance. In Proc. of 2017 International Conference on Image Processing Theory, Tools, and Applications, Nov. 2017. © 2017 IEEE. 3

[17] Robert D. Friedlander and Anthony J. Yezzi. Closed-form solution for thin lens image irradiance under arbitrary solid angle. J. Opt. Soc. Am. A, 37(4):568–578, Apr. 2020. 2, 3, 4

[18] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7675–7684, June 2019. 2

[19] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. CoRR, abs/1906.06543, 2019. 2

[20] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2326–2334, Oct. 2017. 2

[21] Hailin Jin, D. Cremers, D. Wang, E. Prados, Anthony Yezzi, and Stefano Soatto. 3-D reconstruction of shaded objects from multiple images under unknown illumination. Int. Journal of Computer Vision, 76:245–256, Mar. 2008. 2

[22] Hailin Jin, Daniel Cremers, Anthony Yezzi, and Stefano Soatto. Shedding light on stereo segmentation. In Proc. of Computer Vision and Pattern Recognition, pages 36–42, July 2004. 2

[23] Hailin Jin and Paolo Favaro. A variational approach to shape from defocus. In Proc. of the Eur. Conf. on Computer Vision, pages 18–30, 2002. 2

[24] Hailin Jin, Stefano Soatto, and Anthony Yezzi. Multi-view stereo beyond Lambert. In Proc. of Computer Vision and Pattern Recognition, pages 171–178, June 2003. 2

[25] Hailin Jin, Stefano Soatto, and Anthony Yezzi. Multiview stereo reconstruction of dense shape and complex appearance. Int. Journal of Computer Vision, 63:175–189, July 2005. 2

[26] Hailin Jin, Anthony Yezzi, and Stefano Soatto. Variational multiframe stereo in the presence of specular reflections. In Proc. of 3D Data Processing Visualization and Transmission, pages 626–630, June 2002. 2

[27] Hailin Jin, Anthony Yezzi, and Stefano Soatto. Region-based segmentation on evolving surfaces with application to 3D reconstruction of shape and piecewise constant radiance. In Proc. of European Conf. Computer Vision, pages 114–125, May 2004. 2

[28] Hailin Jin, Anthony Yezzi, and Stefano Soatto. Mumford-shah on the move: Region-based segmentation on deforming manifolds with application to 3-D reconstruction of shape and appearance from multiview images. J. Mathematical Imaging and Vision, 29:219–234, Nov. 2007. 2

[29] Hailin Jin, Anthony Yezzi, Yen-Hsi Tsai, Li-Tien Cheng, and Stefano Soatto. Estimation of 3D surface shape and smooth

radiance from 2D images. Journal of Scientific Computing, 19:267–292, Dec. 2003. 2

[30] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. International Journal of Computer Vision, 38(3):199–218, 2000. 2

[31] Feng Li, Jian Sun, Jue Wang, and Jingyi Yu. Dual-focus stereo imaging. *Journal of Electronic Imaging*, 19:043009, 10 2010. 2

[32] Yi Ma, Stefano Soatto, Jana Kosecka, and Shankar S. Sastry. An Invitation to 3D Vision: From Images to Geometric Models, chapter 3. Springer, 2004. 2

[33] Philip F. McLauchlan. Gauge independence in optimization algorithms for 3D vision. In Proc. ICCV '99 Vision Algorithms Workshop, 1999. 2

[34] Kai Ni, Drew Steedly, and Frank Dellaert. Out-of-core bundle adjustment for large-scale 3D reconstruction. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8, Oct. 2007. 2

[35] Stanley Osher. Level set methods. In Geometric Level Set Methods in Imaging, Vision, and Graphics, pages 3–20. Springer-Verlag, New York, 2003. 2

[36] Stanley Osher and James A. Sethian. Fronts propogating with curvature-dependent speed: algorithms based on Hamilton-Jacobi equations. J. of Comp. Physics, 79:12–49, 1988. 2

[37] C. Paramanand and A. N. Rajagopalan. Depth from motion and optical blur with an unscented kalman filter. *IEEE Transactions on Image Processing*, 21(5):2798–2811, 2012. 2

[38] Alex P. Pentland. A new sense for depth of field. IEEE Transactions on Pattern Analysis and Machine Intelligence, 9:523–531, 1987. 2

[39] A. Punnappurath, A. Abuolaim, M. Afifi, and M. S. Brown. Modeling defocus-disparity in dual-pixel sensors. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2020. 2

[40] A. N. Rajagopalan, S. Chaudhuri, and Uma Mudenagudi. Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1521–1525, 2004. 2

[41] Stefano Soatto, Anthony Yezzi, and Hailin Jin. Tales of shape and radiance in multiview stereo. In Proc. of Int. Conf. Computer Vision, pages 974–981, 2003. 2

[42] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. Int J Comput Vis, 128:910–930, Jan. 2020. 2

[43] Murali Subbarao and Gopal Surya. Depth from defocus: A spatial domain approach. International Journal of Computer Vision, 13:271–294, 1994. 2

[44] Y. Takeda, S. Hiura, and K. Sato. Fusing depth from defocus and stereo with coded apertures. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 209–216, 2013. 2

[45] 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, 2013. 2

[46] Gozde Unal, Anthony Yezzi, Stefano Soatto, and Greg Slabaugh. A variational approach to problems in calibration of multiple cameras. Trans. Pattern Anal. Machine Intell., 29:1322–1338, Aug. 2007. 2

[47] T. Wang, M. Srikanth, and R. Ramamoorthi. Depth from semi-calibrated stereo and defocus. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3717–3726, 2016. 2

[48] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images. In 2019 IEEE International Conference on Computer Vision (ICCV), pages 2690–2698, 10 2019. 2

[49] N. Xu and N. Ahuja. On the use of depth-from-focus in 3D object modeling from multiple views. In *Proc. of Asian Conference of Computer Vision*, pages 1038–1043, Jan. 2003. 2

[50] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. Int J Comput Vis, 128:53–73, Aug. 2020. 2

[51] Anthony Yezzi and Stefano Soatto. Stereoscopic segmentation. International J. of Computer Vision, 53:31–43, June 2003. 2, 4, 6

[52] Anthony Yezzi and Stefano Soatto. Structures from motion for scenes without features. In Proc. of Computer Vision and Pattern Recognition, pages 525–532, 2003. 2

[53] Anthony Yezzi, Stefano Soatto, Hailin Jin, Andy Tsai, and Alan Willsky. Mumford-Shah from segmentation to stereo. In Geometric Level Set Methods in Imaging, Vision, and Graphics, pages 207–228. Springer-Verlag, New York, 2003. 2