Finite Aperture Stereo: 3D Reconstruction of Macro-Scale Scenes

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

Abstract

While the accuracy of multi-view stereo (MVS) has continued to advance, its performance reconstructing challenging scenes from images with a limited depth of field is generally poor. Typical implementations assume a pinhole camera model, and therefore treat defocused regions as a source of outlier. In this paper, we address these limitations by instead modelling the camera as a thick lens. Doing so allows us to exploit the complementary nature of stereo and defocus information, and overcome constraints imposed by traditional MVS methods. Using our novel reconstruction framework, we recover complete 3D models of complex macro-scale scenes. Our approach demonstrates robustness to view-dependent materials, and outperforms state-of-the-art MVS and depth from defocus across a range of real and synthetic datasets.

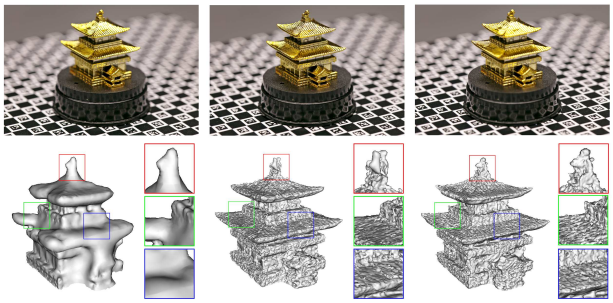

Figure 1. Top row: Example focal stack input images used by our approach from our real-world 16 view Temple dataset. Bottom row: This object is very challenging for MVS [54] (left), multi-view DFD performs much better [3] (middle) but the proposed achieves the best result (right). Notice the better recovery of the roof ornament, sharper edges and improved surface details.

1. Introduction

Passive scene reconstruction continues to be an actively researched problem. Often, a multi-view stereo approach is taken to overcome occlusion and achieve complete scene modelling. Despite the advances made in conventional works and the recent adoption of deep learning, the performance of MVS remains heavily dependent on the scene content. Accurate correspondence between viewpoints is only possible when surfaces are uniquely textured and consistent in appearance. As a result, materials with periodic textures or complex light interactions such as sub-surface scattering cannot be reconstructed without heavy reliance on handcrafted or learnt scene priors.

An additional limitation, often overlooked in literature, is the simplification of the image formation process. This pinhole camera model traditionally used in MVS only considers the scene projection, and assumes images are free from all optical aberrations. While some aberrations such as lens distortion can be detected during calibration and corrected, defocus as a result of a finite aperture cannot. As a result, the scene is implicitly limited to the depth of field (DoF) of the camera where pixels can be considered acceptably sharp. Consequently, macro-level reconstructions where the DoF is shallow are not theoretically possible.

From this perspective, defocus can usually be considered an undesirable artifact. However, given accurate camera parameters, its formation on the image plane can be modelled from simple geometric principles and related to scene depth. The exploitation of this phenomena for scene reconstruction is known as depth from defocus (DFD), and is a well estabilished area of research. Since focus analysis is monocular in nature, DFD techniques are suitable for recovering complex materials and textures that would otherwise be challenging for multi-view stereo. However, traditional implementations only achieve partial reconstructions because additional views are not used.

Considering these complementary properties, it is clearly advantageous to combine these two approaches into a unified framework. This reasoning is consistent with biological studies as noted in [26], where defocus information is shown to improve stereo matching in human vision. While some previous works have explored this idea, no work that we are aware of has performed a study in a multi-view

Published in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 2474–2484, 2021. DOI: https://doi.org/10.1109/ICCVW54120.2021.00280

context to recover complete 3D models. In this paper, we present exactly this, and demonstrate the benefits of combining an accurate defocus camera model with standard MVS projective geometry principles.

Our proposed MRF-based framework combines both cues while retaining their individual advantages - the convexity and stability of defocus vs the enforced geometric consistency and high reconstruction fidelity of MVS. As input, we take posed multi-view focal stacks which focal-sweep the volume of interest from relatively sparse viewpoints. The limited availability of this type of data makes the application of a deep learned-based approach difficult, and motivates our more traditional methodology. We show how stereo information can be applied to finite aperture images with significant defocus without violating traditional pinhole assumptions. Surprisingly, our approach indicates stereo correspondence can improve reconstruction even in the presence of view-dependent materials (see Figure 1).

Unlike previous works, we avoid iterative estimation of defocus parameters by modelling the camera as a thick lens [3]. This defocus model is well suited for combining with multi-view observations, since the blurring response is explicitly calibrated relative to the projective center of the camera; making for elegant integration into our framework. In contrast to [3], who achieve 3D reconstructions using only defocus cues, we demonstrate how including stereo information helps recover higher accuracy geometry.

In our evaluation we compare performance to state-of-the-art MVS and DFD on novel multi-view datasets, and achieve superior results reconstructing scenes containing specular and reflective surfaces. In summary, this paper presents the following:

1. A framework unifying a thick-lens defocus model with multi-view stereo
2. An iterative algorithm to overcome the limitations a narrow DoF imposes on reconstruction fidelity
3. Comparative evaluation of performance on complex datasets to state-of-the-art MVS and DFD

The remainder of this paper is structured as follows. Section 2 discusses previous work. Section 3 briefly analyses the formation models of each cue. Section 4 explains the proposed approach, and section 5 performs a comparative evaluation. Section 6 concludes the paper.

2. Previous Work

In this section, we survey related work. Here, stereo-based and focus-based reconstruction approaches are covered, and we include works considering these cues individually or in combintation. To clarify the often interchanged terminology used in focus-based reconstruction and to keep the survey concise, we largely exclude approaches which evaluate the structure of a scene from a focal stack based on the response of a focus measure e.g. [46].

2.1. Multi-View Stereo

Perhaps one of the most widely understood reconstruction principles, MVS recovers 3D structure by identifying corresponding features from images of the scene taken at different viewpoints. Using geometric constraints arising from the pinhole camera model, 3D points can be triangulated from two or more of these features according to the pose of each view. Broadly speaking, the quality of reconstruction largely depends on three factors.

Scene Representation: How surfaces are modelled not only affects the resolution of the final result, but also places restrictions on the reconstruction algorithm. For instance, voxel-based [63, 41, 27, 30, 14] and mesh-based [37, 15] representations allow for a globally optimal result, since all views can be evaluated jointly. Alternatively, view-dependent methods [54, 62] only use a subset of the input images to recover a depth map of each viewpoint.While they do not impose the strict initialisation of voxel-based and mesh-based methods, they require post processing and produce potentially less robust results.

Feature Matching: At the heart of all MVS algorithms is a similarity metric used to identify corresponding points between images. Classical metrics implement per-pixel comparisons such as sum of squared differences (SSD) [36] and normalised cross correlation (NCC) [37, 7, 21]. Some works exploit perspective distortion to also estimate surface normals [7]. More recent approaches generally use feature descriptors to extract richer information from the source images. Though initially hand-crafted [61, 62], the advent of deep learning introduced data-driven feature extraction with convolutional neural networks (CNNs) [69, 67].

Regularisation: To overcome the real-world limitations of standard MVS assumptions, most approaches use a regularisation framework to enforce scene priors. A popular traditional approach involves formulating these priors as part of an energy function, and solving with a Markov Random Field (MRF). Early deep learning works followed a similar idea, though recent approaches regularise with learnt priors.

Of particular interest to this survey is view-dependent methods. Many conventional approaches were able to produce compelling results despite the limitations of traditional feature matching, often resulting in creative methodologies [70, 40, 62]. Notably, PMVS [21] combine matched patches rather than point clouds, and refine the final mesh using an energy optimisation to impose smoothness constraints. COLMAP [54], arguably one of the best performing conventional MVS methods, combines a structure from motion calibration with a view dependent reconstruction pipeline to produce high quality 3D models.

More recently, deep learning-based approaches have seen widespread success. SurfaceNet [29] introduced the first method trained end-to-end based around a voxel grid. DeepMVS [28] instead generates a plane sweep volume and aggregates matched features from an arbitary number of images. MVSNet [67] introduces differentiable homography warping, and R-MVSNet [68] improves the memory efficency with a recurrent architecture. PointMVSNet [11] adopts a coarse-to-fine approach with multi-scale features. CasMVSNet [24] develops a memory efficent cost volume and adapts it to existing methods. Though not advertised as MVS, neural radiance fields [45] achieve dense implicit reconstructions. Other notable works include [42, 34].

2.2. Depth from Defocus

By modelling the point spread function (PSF) of the camera, depth information of the scene can be leveraged from the formation of defocus on the image plane. DFD is a field of research that approaches this idea in many different and creative ways. Though techniques exist for evaluating depth from a single defocused image [9, 2, 8, 31], we primarily focus on methods that require several defocused images captured with circular apertures.

Acquisition: A convenient method for capturing multiple defocused images is with a lightfield camera [59]. However, lightfield cameras can only capture the scene at a limited resolution. With conventional camera lenses, there are two main approaches to generate differently focused images - with varying aperture size [50, 44, 55] or focusing distance [20, 47]. Changing the aperture size is often simpler, but the scene reconstruction volume is limited due to the relative blur exhibiting a symmetrical transfer function [43]. Although focal stacks largely overcome this ambiguity, refocusing the camera in this way introduces scale and translational differences between images and subsequently requires correcting [65, 58, 4, 3]. Some methods [25] vary both the aperture size and focus setting to capture dense information about the camera PSF.

PSF Modelling: Most approaches assume a convolutional formation model, allowing the PSF to be approximated as a 2D kernel. Two popular choices include the Pillbox [65, 18] and Gaussian [20, 4, 51] functions. These methods do not consider many of the abberations present in optical systems, so some works [31, 44] instead directly measure the blurring response of the camera. Other works do not model the PSF explicitly, instead depending on a data driven approach [25, 8, 19]. In many cases, a thin lens defocus model is assumed despite the fact this model does not hold in real-world optical systems. [39] improves reconstruction accuracy through iterative refinement. [49] considers a model beyond a thin lens, and formulates sub-aperture disparity relative to the entrance pupil in a colour coded-aperture camera. [3] proposes a formal calibration of a thick lens camera model, and applies it to capturing and reconstructing multi-view focal stacks.

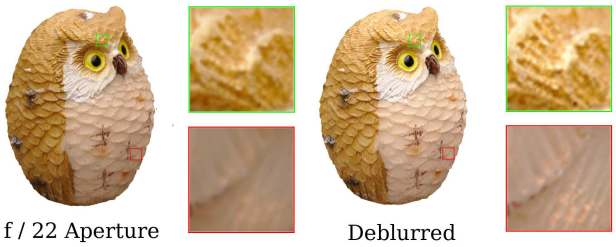

Figure 2. Difference in image quality between a pinhole aperture (left) and a deblurred focal stack generated by our method (right). The small aperture image is degraded by diffraction blurring arising from the wave-like nature of light. This is not the case with larger apertures, capturing brighter and more detailed surfaces.

Aside from [16] who utilise deep learning, most works adopt an MRF-based or numerical optimisation framework. Moreover, the overwhelming majority of DFD methods discussed only achieve single-view reconstructions. This is in part due to limitions modelling the PSF, as well as a lack of publically available datasets. To our knowledge, [3] is the only attempt at 3D reconstruction using only defocus cues; by fusing multiple single-view reconstructions together.

2.3. Hybrid Approaches

We will now discuss previous works that take advantage of multiple reconstruction cues. Most existing methods formulate their combination of stereo and defocus in an MRF framework. One approach is to combine cues with defocused stereo pairs [35, 52, 10]; often expressing the relative blurring kernel in terms of pixel disparity. [57] apply coded apertures in this way. [1] instead uses defocus to constrain stereo matching. Other methods apply single-image defocus constraints to better recover discontinuities [64, 23].

Alternative to pairwise-stereo, some methods use lightfield cameras to combine cues [38, 59, 60], though reconstructions are limited to a very narrow baseline. [5] consider multiple viewpoints, but do not apply this to 3D reconstruction. [12] is the only approach we know of to use deep learning for combining cues. However, as with all works discussed, reconstructions remain limited to a single view.

Finally, shading cues have been proposed in combination with defocus [13], stereo [66] and both [60] to alleviate the texture requirements of these cues.

2.4. Summary

Though many works have proposed methodologies considering stereo and defocus separately, far fewer have attempted combining them. Those who have limit reconstruction to a single view, foregoing complete scene modelling. In this paper, we fill this gap and evaluate the benefits of these cues in a multi-view context.

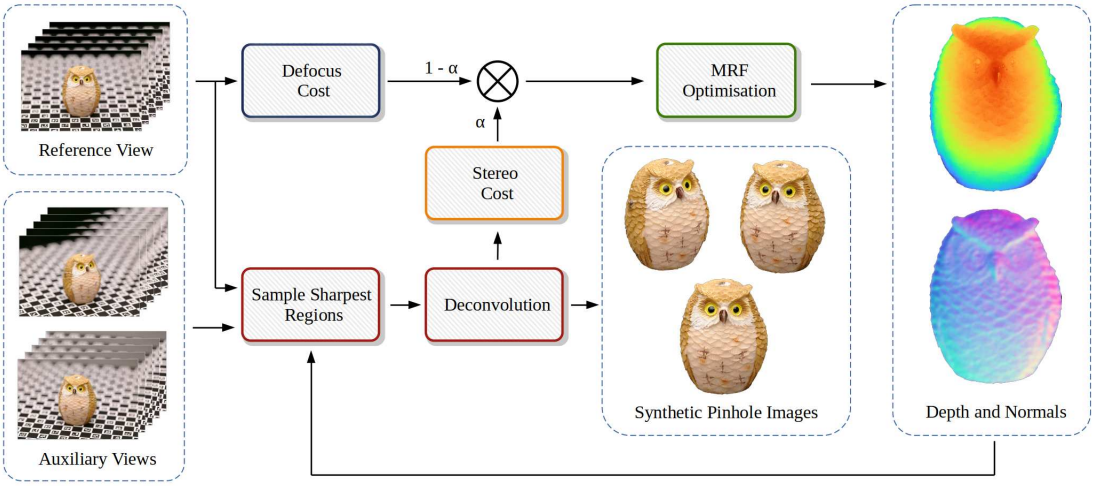

Figure 3. Overview of our approach illustrating the iterative nature of the proposed pipeline. Defocus and stereo costs are generated from the calibrated focal stacks and synthetically deblurred images respectively, then combined according to the value of α. This weighted sum is input to an MRF framework, where spatial consistency is enforced according to second order smoothness priors. The output from the MRF is the estimated depth, which is used in the next iteration to re-generate pinhole images of the focal stacks. As iteration increases, α is updated and the effective resolution of the pipeline doubles. This process continues until the maximum number of iterations has been reached. To generate 3D models, the depth and normal maps from each viewpoint are converted to point clouds, and merged together.

3. Image Formation

Though stereo and defocus cues operate on very different principles, the image formation assumptions made by each cue can be generalised easily; allowing for a brief analysis of their differences. [53] provide a more in-depth analysis. For simplicity, we ignore the effects of lens distortion as this can be corrected computationally. The formation of a pixel $\boldsymbol{y}$ on image I can be described [20]

$$I(\boldsymbol{y}) = \int k(\boldsymbol{y}, \boldsymbol{x})\, r(\boldsymbol{x})\, d\boldsymbol{x}, \tag{1}$$

where $r(\boldsymbol{x})$ defines the radiance of the projected world coordinates $\boldsymbol{x}$ and k describes the PSF of the camera. Given the same scene and camera pose, the only differences in this context between cues is how the camera response k is modelled. As previously discussed, a pinhole camera is typically assumed in most MVS methods. In this case, k is equal to a Dirac delta and the projected image represents the incident radiance. However, in reality, small apertures give rise to diffraction blurring and degrade the overall sharpness of the image. An example of this can be seen in Figure 2.

Defocus models instead take into account a finite aperture. Usually, Equation 1 is approximated as a convolution and k is modelled as a kernel that best represents the aperture shape. The size of the blurring kernel σ is related geometrically to scene depth relative to the camera pinhole d,

$$\sigma(d) = \frac{\gamma a v}{2}\left(\frac{1}{d-w} + \frac{1}{v} - \frac{1}{f}\right) \tag{2}$$

according to a constant γ, aperture a, focus setting v, focal length f and pupillary magnification offset w as defined by the thick lens defocus model. For a thin lens model, $w = 0$. We refer the reader to [3] for further details.

From the above, it is clear neither cue models the light reflected from a scene point beyond a simple projective transform. In other words, the light transport of the scene is not considered prior to the final surface interaction. While this is detrimental to MVS in the presence of view-dependent materials, defocus information remains coherent due to its monocular nature. Since defocus is a camera-centric phenomena, the reconstruction principles of DFD can be generalised across many complex scenes with little regard to their content provided sufficent defocus-variant texture is present [17]. At the macro-scale magnification explored in this paper, this limitation is not a concern.

4. Methodology

Our approach combines defocus and stereo information to leverage the benefits of both cues to generate complete 3D models of macro-scale scenes. The proposed pipeline can be broken up into two sequential stages.

Reconstruction: Using stereo and defocus cues, we reconstruct per-viewpoint depth maps as shown in Figure 3. As input, we take multi-view focal stacks captured and calibrated using the approach proposed in [3]. These images have a narrow DoF, making them unsuitable for direct stereo matching. As part of our pipeline, we deblur these focal stacks via non-blind deconvolution, and perform matching on the synthetically sharpened images. The two cues are then jointly optimised to find the surface estimate, which is

refined in subsequent iterations. Our approach can be summarised as follows:

1. Calculate an initial thick-lens DFD reconstruction
2. Deblur the input focal stacks using the camera model and estimated depth to approximate scene radiance
3. Find corresponding points from synthesised radiance
4. Combine defocus and correspondence information and recalculate surface at higher resolution
5. Repeat steps 2, 3 and 4 until maximum resolution or iteration reached

Point Cloud Fusion: The point clouds from each view are combined to produce the final 3D model. We enforce consistency checks on each reconstructed point to reduce noise, before applying screened Poisson surface reconstruction [32] to generate the final triangular surface mesh.

4.1. Energy Function

As in [3], we formulate depth recovery of each view as a discrete labelling problem of N labels, which we generalise here to exploit both defocus and stereo cues. Each cue is represented as a data term in our energy function,

$$E(\boldsymbol{x}, n) = (1 - \alpha(n)) \sum_{p \in \nu} \Phi_D(x_p) + \alpha(n) \sum_{p \in \nu} \Phi_S(x_p) \; + \frac{\lambda}{2^{n-1}} \sum_{(p,q) \in \epsilon} \Psi_{pq}(x_p, x_q). \tag{3}$$

Here, α is a scalar value between 0 and 1, and weights the contributions of the defocus term Φ_D and the stereo term Φ_S. We linearly modulate its value with increasing iteration up to a maximum of 0.5. The value of λ controls the amount of pairwise smoothness applied by Ψ_{pq}, which encourages second order smoothness as described in [48].

In our framework, we assume each pixel represents a surface and model it as a tangent plane. During reconstruction, the candidate search space of each surface is independently reduced as a function of iteration n. Unlike traditional MRF formulations, this approach allows for high resolution reconstructions without requiring a corresponding number of labels; reducing memory usage and computational load. As n increases, the effect of the smoothness term is decreased to enable the recovery of higher fidelity surface details. Equation 3 is minimised using α-expansion [6, 56]. We will now explain each of the terms in Equation 3, with particular emphasis on the novel stereo term.

Figure 4. Materials simulated in our synthetic datasets: stone Armadillo (left), gold Bunny (middle) and wooden Dragon (right).

4.2. Defocus Term

To integrate defocus information into our framework, we implement the defocus term defined in [3]. First, harmonic texture components are removed from the focal stacks [17]. By assuming a convolutional image formation model, the defocus observed on the image plane as a function of scene depth d can be described by Equation 2. From this, the defocus term determines the relative blur between pairs of images in the reference view focal stack. For focus settings i and j, this relative blur is defined

$$\sigma_{ij}(d) = \sqrt{|\sigma_i(d)^2 - \sigma_j(d)^2|}. \tag{4}$$

Given a candidate label depth d_k, the defocus term generates a photometric cost by blurring the sharper image to match the other according to Equation 4. We denote this blurring according to a defocus operator $\circ$,

$$\phi_D(x_{\boldsymbol{p}}) = \sum_{\{ij\} \in \boldsymbol{\Omega_D}} \sum_k \left(\sigma_{ij}(d_k) \circ I_a \; - \; I_b \right)^2, \tag{5}$$

$$\{a, b\} \; = \; \begin{cases} \{i, j\} & \sigma_i(d) < \sigma_j(d), \\ \{j, i\} & \text{otherwise.} \end{cases} \tag{6}$$

Here, I is an input image and $\boldsymbol{\Omega_D}$ defines the set of image pairs. In practise, $\circ$ models defocus as a diffusion process which is equivalent to a Gaussian PSF. Further details of this term can be found in [3]. Finally, the generated cost volume is normalised according to

$$\Phi_D(x_{\boldsymbol{p}}) \; = \; 1 - \exp\left(-\frac{\phi_D(x_{\boldsymbol{p}})}{\mu_D}\right), \tag{7}$$

where μ_D is the mean of ϕ_D.

4.3. Stereo Term

While the defocus term has a stable response in the presence of defocus-variant texture, it does not necessarily permit the recovery of high frequency surface detail. This is a consequence of the nature of defocus blur; surface details are attenuated by the aggregation of photons in out-of-focus regions. The stereo term is intended to improve the fidelity of the reconstruction by integrating correspondence information from synthetically deblurred images.

To this end, we deblur the input focal stacks according to the current surface estimate. This is achieved via non-blind deconvolution using a Wiener filter. Patches are then

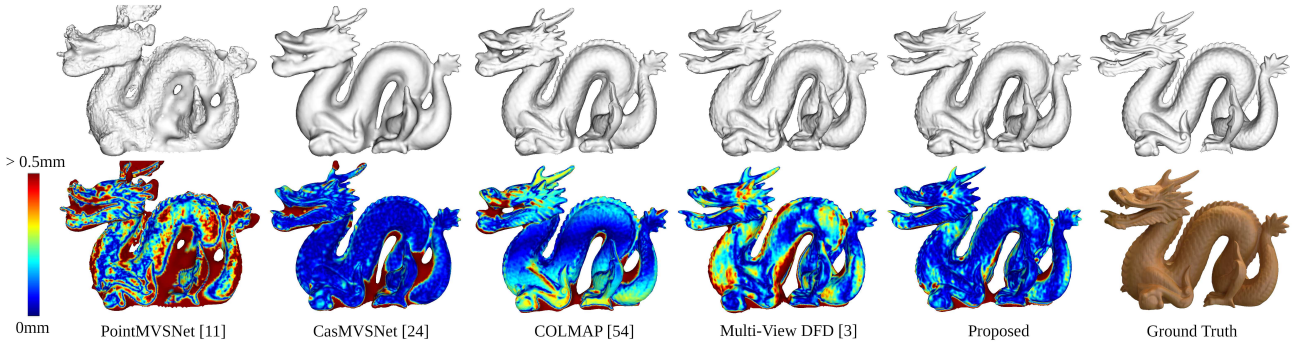

Figure 5. Top row: Poisson surface reconstruction results on the synthetic wooden Dragon dataset for each method tested. Bottom row: error maps when compared to the ground truth mesh ranging from 0mm (blue) to 0.5mm (red) error.

extracted from auxillary views taking into account the orientation of the surface, and compared to the reference view. In our implementation, one view from either side of the reference view is used to improve robustness to occlusions. Let us now look in detail at how a single pixel p is processed.

4.3.1 Focal Stack Deblurring

For each iteration, the input focal stacks are deblurred according to the current surface estimate. This is the case not only for the reference view, but the auxillary viewpoints as well. For pixel p, the first step is to determine its depth d^p by raytracing and intersecting the estimated surface. Using the thick lens camera model, we then determine which image from the focal stack $\boldsymbol{I}$ exhibits the least amount of blur at p by minimising Equation 2 for all images. To prevent the filter from becoming unstable where the surface estimate is inaccurate or resolution too coarse, this value is divided by 2 and truncated to a maximum value. We denote the minimum blur of pixel p as σ^p.

Next, a Wiener filter is utilised to implement a spatially variant deconvolution. A PSF kernel of size σ^p is generated, and its corresponding Fourier transform $K_\sigma(\omega)$ calculated. The deconvolution kernel in the frequency domain $G(\omega)$ necessary to sharpen p is

$$G(\omega) = \frac{K^*_\sigma(\omega)}{|K_\sigma(\omega)|^2 + \epsilon}, \tag{8}$$

where $*$ represents complex conjugation, and ϵ is a constant representing the inverse signal-to-noise ratio (SNR). By setting this to a small value, we implicitly assume a relatively noise-free image. This is not an unreasonable assumption given the large aperture used during the capture of the focal stacks. A square patch w_p is then extracted from the image where p appears sharpest. The size of w_p is determined according to σ^p. Finally, the deblurred pixel $\hat{p}$ is calculated

$$\hat{p} = \left|\left| \mathcal{F}^{-1}\Big(W_p(\omega).G(\omega)\Big) \right|\right|, \tag{9}$$

with $\mathcal{F}$ denoting a Fourier transform and $W_p(\omega) = \mathcal{F}(w_p)$. This spatially variant deconvolution is performed for all valid pixels in all relevant input views.

4.3.2 Patch Matching

Now the focal stacks have been deblurred, correspondence information can be acquired. Assuming p is in the reference view, we define a square support patch w_p centred around p, and cast rays through every pixel within it into world-space. The ray corresponding to p is sampled along according to the candidate labels. The remaining rays are intersected with planes centred at each sample with normals equal to that of the surface. When projected into the auxillary views, the shape of the resulting support patch will distort to better resemble the appearance of the surface in those views. To account for subpixel sampling arising from this projection, we implement bilinear interpolation. Our matching cost between the patch in the reference view w_p and a patch in the auxillary view w_q is defined by the pixel-wise comparison

$$\phi_S(x_{\boldsymbol{p}}) = \sum_{\{j\} \in \boldsymbol{\Omega_S}} \sum_{p} \left(w_p - w_p^j\right)^2, \tag{10}$$

where $\boldsymbol{\Omega_S}$ defines a vector of auxillary views. Similar to the defocus term, Equation 10 is normalised to produce the final stereo term, where μ_S is the mean of ϕ_S:

$$\Phi_S(x_{\boldsymbol{p}}) \;=\; 1 - \exp\left(-\frac{\phi_S(x_{\boldsymbol{p}})}{\mu_S}\right). \tag{11}$$

4.4. Smoothness Term

The purpose of the smoothness term is to ensure the reconstructions remain coherent in textureless or saturated regions while retaining surface edges. The general form of such a function can be written [56]

$$\Psi_{pq}(x_p, x_q) = \min\left(\Psi_{max}, V_{pq}(x_p, x_q)\right). \tag{12}$$

The above enforces pairwise smoothness between two pixels p and q taking labels x_p and x_q respectively, with the

		Armadillo			Bunny			Dragon		
		Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood
COLMAP [54]	0%	0.9289	0.9904	**0.9893**	0.7996	0.9578	**0.9585**	0.7819	0.9186	0.9193
	1%	**0.9285**	**0.9906**	**0.9872**	0.7900	**0.9566**	**0.9553**	0.7724	**0.9162**	**0.9150**
CasMVSNet [24]	0%	0.9093	0.9645	0.9642	0.8621	0.9462	0.9456	0.8313	0.9010	0.9017
	1%	0.9128	0.9624	0.9636	**0.8518**	0.9505	0.9463	0.8270	0.8990	0.9007
PointMVSNet [11]	0%	0.7778	0.8759	0.8894	0.7853	0.8962	0.8939	0.7542	0.8374	0.8634
	1%	0.7964	0.8919	0.9159	0.8011	0.8883	0.8942	0.7609	0.8421	0.8684
Multi-View DFD [3]	0%	**0.9880**	0.9888	0.9755	**0.9411**	0.9502	0.9374	**0.9276**	0.9346	0.9276
	1%	0.8903	0.7224	0.6420	0.7908	0.7147	0.6681	0.7791	0.6520	0.6071
Proposed	0%	0.9541	**0.9906**	0.9887	0.8740	**0.9590**	0.9536	0.8893	**0.9494**	**0.9454**
	1%	0.9278	0.9265	0.9276	0.8028	0.8852	0.8871	**0.8488**	0.8503	0.8792

Table 1. F-score results (τ = 1mm) on point cloud outputs for our synthetic dataset with 0% and 1% additive Gaussian noise. Bold indicates top performer for each material and noise level.

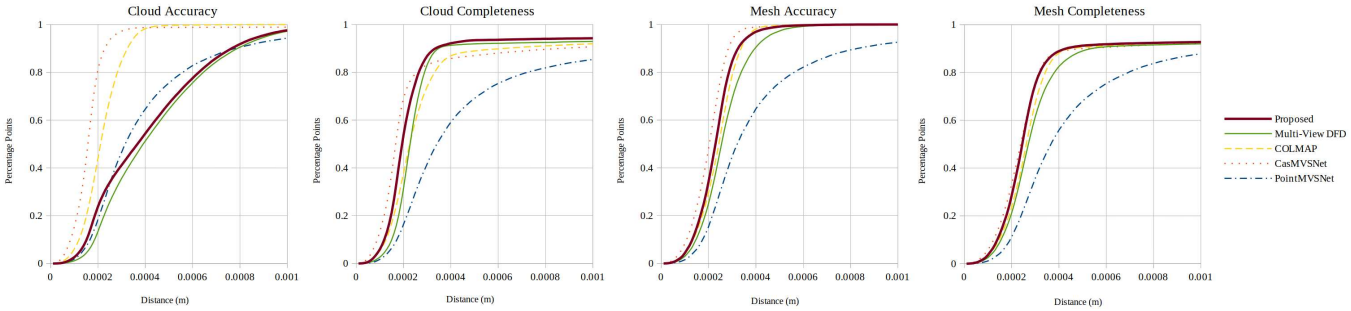

Figure 6. Point cloud accuracy and completeness histograms on the synthetic stone Bunny dataset (left). Since our method produces many more points than the MVS methods, our apparent accuracy suffers due to normalisation. We therefore present the same analysis on the Poisson meshes (right) where the resolution of vertices is more consistent.

truncation preserving discontinuities. Following [3], we define V_{pq} as a second-order prior and exploit the tangent plane surface model. For two world-points $\boldsymbol{P}$ and $\boldsymbol{Q}$ corresponding to labels x_p and x_q respectively, we define V_{pq}

$$V_{pq}(x_p, x_q) = \left(\frac{1}{\delta(n)(N-1)} \left| \frac{(\boldsymbol{Q}-\boldsymbol{P}) \cdot \boldsymbol{q^n}}{\boldsymbol{p^r} \cdot \boldsymbol{q^n}} \right| \right)^2, \quad (13)$$

similar to the definition proposed in [48]. Here, $\boldsymbol{q^n}$ is the normal of the surface related to pixel q, $\boldsymbol{p^r}$ is a ray cast through pixel p and $\delta(n)$ is the metric distance between labels. This expression penalises label assignment based on the curvature of the surface, enabling a smooth piece-wise linear reconstruction. In our framework, we set $\Psi_{max} = 0.1$ and $\lambda = 10000$.

4.5. Point Cloud Fusion

To filter out significantly erroneous points in the point cloud outputs, a post-processing correspondence check is performed. Our implementation requires each point to correspond in at least two adjacent views to within 1mm. We also exclude corresponding points where the difference in normal vectors exceeds 30 degrees. All remaining points are then subject to screened Poisson surface reconstruction to generate a complete mesh of the scene.

5. Evaluation

To evaluate our approach, we compare performance on synthetic and real data to three view-dependent MVS approaches; PointMVSNet [11], CasMVSNet [24] and COLMAP [54]. Instead of operating on focal stacks, these methods take pinhole images as input. When operating on real data, these pinhole images are captured with an f/22 aperture (see Figure 2). All synthetic and real inputs are 16-bit images with a resolution of 2184x1464 pixels, although the MVS methods require 8-bit input images instead.

PointMVSNet and CasMVSNet were run pre-trained with 128 labels on an Nvidia RTX 3070 with 8GB of VRAM, and point clouds were combined using code from [22] as instructed by the authors. Due to memory restrictions, the input images were downsampled to a quarter resolution. Otherwise, parameters were left largely at their default values. We also compare to a re-implementation of multi-view DFD [3], by setting $\alpha = 0$ in Equation 3. In all cases, our approach and DFD use a visual hull initialisation for the first iteration, but is then disabled for all subsequent iterations. The MVS methods do not use this visual hull information. Aside from this, all calibraton information and auxillary views are kept consistent between methods.

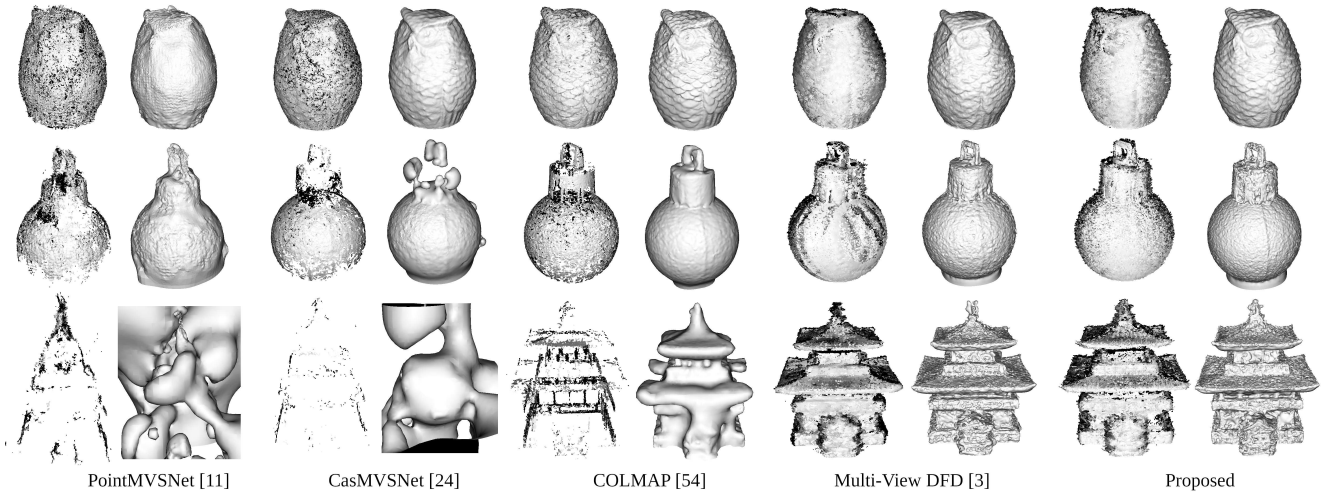

Figure 7. Point cloud (left columns) and Poisson surface mesh (right columns) results from each method on real datasets. Here, we show reconstructions on the Owl dataset from [3] (top row), and the datasets we captured - Bauble (middle row) and Temple (bottom row). The many specularities present in our datasets makes photometric consistency difficult to determine, resulting in sparse point clouds from the MVS methods. In all cases, the proposed approach recovers stable geometry and detailed surfaces.

5.1. Synthetic Data

We generated several multi-view synthetic datasets using the Stanford Armadillo, Bunny and Dragon to verify our approach. These datasets consist of 24 views, and simulate several challenging materials as seen in Figure 4. Gaussian noise has been added to better reflect real-world conditions. For DFD and the proposed approach, we generate 5-image focal stacks for each viewpoint with a narrow DoF to simulate a finite aperture. Figure 5 qualitatively demonstrates the difference in performance on the Dragon. Quantitive evaluation was performed by comparing the reconstructed point clouds with the ground truth geometry, using the F-score metric proposed by [33]. Table 1 presents these results. In many cases, the proposed approach outperforms MVS and DFD, and otherwise performs competitively.

Figure 6 presents histograms showing accuracy and completeness for each method on the stone Bunny dataset. We present results on the point clouds and Poisson meshes to better understand the performance of each approach. Our method improves accuracy over DFD alone ($\alpha = 0$), and achieves a high degree of overall completeness. This is particularly apparent in Figure 5 when generating the final surface mesh, showing comparitively fewer significant errors.

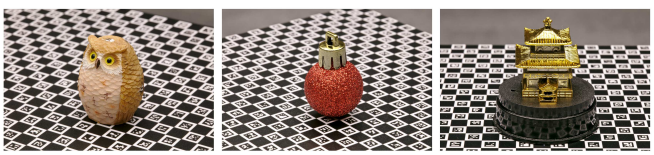

Figure 8. Selected pinhole (aperture f/22) images from Owl (left), Bauble (middle) and Temple (right) datasets.

5.2. Real Data

To compare these methods on real data, we captured two multi-view, multi-focus datasets using the technique proposed in [3]: Bauble (18 views) and Temple (16 views). These objects exhibit view-dependent properties, making them difficult to reconstruct using conventional methods. A qualitative comparison of performance on these datasets can be seen in Figure 7, with their appearance in Figure 8. Our approach produces stable and complete reconstructions; recovering details absent from the other methods. We also compare performance on the Owl dataset from [3], which is better suited for MVS. The proposed outperforms DFD, and performs as well as or better than the MVS approaches.

6. Conclusion

In this paper, we have developed a framework that unifies the benefits of stereo and defocus information to better recover geometry from finite aperture images. We proposed a novel method to overcome the traditional limitations of MVS, and experimentally proven its effectiveness across many synthetic and real materials. Though our stereo term alone would struggle in comparison to the advanced MVS algorithms we have compared against, in combination with defocus information it produces compelling results on challenging macro-scale scenes. In future work, we intend to explore how to combine these cues in a spatially-variant manner, to improve performance in the presence of occlusion and signficant non-Lambertian surfaces.

Acknowledgments This research was supported by the EPSRC (grants EP/N509772/1, EP/P022529/1).

References

[1] Arnav Acharyya, Dustin Hudson, Ka Wai Chen, Tianjia Feng, Chih-Yin Kan, and Truong Nguyen. Depth estimation from focus and disparity. volume 2016-, pages 3444–3448. IEEE, 2016. 3

[2] Saeed Anwar, Zeeshan Hayder, and Fatih Porikli. Deblur and deep depth from single defocus image. *Machine vision and applications*, 32(1), 2021. 3

[3] Matthew Bailey and Jean-Yves. Guillemaut. A Novel Depth from Defocus Framework Based on a Thick Lens Camera Model. In *2020 International Conference on 3D Vision (3DV)*, pages 1206–1215, 2020. 1, 2, 3, 4, 5, 7, 8

[4] Rami 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

[5] Arnav Bhavsar and A. Rajagopalan. Towards unrestrained depth inference with coherent occlusion filling. *International Journal of Computer Vision*, 97(2):167–190, 2012. 3

[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. 5

[7] D Bradley, T Boubekeur, and W Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[8] 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 *Computer Vision – ECCV 2018 Workshops*, pages 307–323. Springer International Publishing, 2019. 3

[9] Ayan Chakrabarti and Todd Zickler. Depth and Deblurring from a Spectrally-Varying Depth-of-Field. In *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 648–661. Springer Berlin Heidelberg, Berlin, Heidelberg. 3

[10] Ching-Hui Chen, Hui Zhou, and Timo Ahonen. Blur-Aware Disparity Estimation from Defocus Stereo Images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 2015, pages 855–863, 2015. 3

[11] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-Based Multi-View Stereo Network. *CoRR*, abs/1908.04422, 2019. 3, 7

[12] Zhang Chen, Xinqing Guo, Siyuan Li, Xuan Cao, and Jingyi Yu. A Learning-based Framework for Hybrid Depth-from-Defocus and Stereo Matching. *arXiv e-prints*, page arXiv:1708.00583, Aug. 2017. 3

[13] Yasuyuki Chen Li, Stephen Shuochen Su, Stephen Matsushita, Stephen Kun Zhou, and Stephen Lin. Bayesian depth-from-defocus with shading constraints. *Image Processing, IEEE Transactions on*, 25(2):589–600, 2016. 3

[14] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In *Computer Vision – ECCV 2016*, pages 628–644. Springer International Publishing, 2016. 2

[15] Amaël Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1493, 2014. 2

[16] David R. Emerson and Lauren A. Christopher. 3-D Scene Reconstruction Using Depth from Defocus and Deep Learning. In *2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8, 2019. 3

[17] P. Favaro. Shape from focus and defocus: Convexity, quasi-convexity and defocus-invariant textures. pages 1–7. IEEE, 2007. 4, 5

[18] Paolo Favaro. Recovering thin structures via nonlocal-means regularization with application to depth from defocus. pages 1133–1140. IEEE Publishing, 2010. 3

[19] 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. 3

[20] P. Favaro, S. Soatto, M. Burger, and S.J. Osher. Shape from defocus via diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):518–531, 2008. 3, 4

[21] Yasutaka Furukawa and Jean Ponce. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 32(8):1362–1376, 2010. 2

[22] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 873–881, 2015. 7

[23] Ioana Gheţa, Christian Frese, Michael Heizmann, and Jürgen Beyerer. A New Approach for Estimating Depth by Fusing Stereo and Defocus Information. In Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, and Marc Ronthaler, editors, *Informatik 2007 – Informatik trifft Logistik – Band 1*, pages 26–31, Bonn, 2007. Gesellschaft für Informatik e. V. 3

[24] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. 2019. 3, 7

[25] Samuel Hasinoff and Kiriakos Kutulakos. Confocal stereo. *International Journal of Computer Vision*, 81(1):82–104, 2009. 3

[26] David M Hoffman and Martin S Banks. Focus information is used to interpret binocular images. *Journal of vision*, 10(5):13,13, 2010-05-01. 1

[27] A. Hornung and L. Kobbelt. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 503–510, 2006. 2

[28] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning Multi-view Stereopsis. *CoRR*, abs/1804.00650, 2018. 3

[29] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2307–2315, 2017. 3

[30] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine, 2017. 2

[31] Masako Kashiwagi, Nao Mishima, Tatsuo Kozakaya, and Shinsaku Hiura. Deep Depth From Aberration Map. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4069–4078, 2019. 3

[32] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM transactions on graphics*, 32(3):1–13, 2013. 5

[33] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 8

[34] Andreas Kuhn, Christian Sormann, Mattia Rossi, Oliver Erdler, and Friedrich Fraundorfer. Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction, 2020. 3

[35] Feng Li, Jian Sun, Jue Wang, and Jingyi Yu. Dual-focus stereo imaging. *Journal Of Electronic Imaging*, 19(4), 2010. 3

[36] Gang Li and S.W Zucker. Differential geometric inference in surface stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):72–86, 2010. 2

[37] Zhaoxin Li, Kuanquan Wang, Wangmeng Zuo, Deyu Meng, and Lei Zhang. Detail-preserving and content-aware variational multi-view stereo reconstruction. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 25(2), 2016. 2

[38] Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. Depth Recovery from Light Field Using Focal Stack Symmetry. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3451–3459, 2015. 3

[39] Xing Lin, Jinli Suo, Xun Cao, and Qionghai 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. 3

[40] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):407–418, 2010. 2

[41] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1052–1061, 2019. 2

[42] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-MVSNet: Learning Patch-Wise Matching Confidence Aggregation for Multi-View Stereo. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10451–10460, 2019. 3

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

[44] Manuel Martinello, Andrew Wajs, Shuxue Quan, Hank Lee, Chien Lim, Taekun Woo, Wonho Lee, Sang-Sik Kim, and David Lee. Dual Aperture Photography: Image and Depth from a Mobile Camera. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1,10, 2015-04. 3

[45] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 3

[46] Michael Moeller, Martin Benning, Carola Schonlieb, and Daniel Cremers. Variational depth from focus reconstruction. *Image Processing, IEEE Transactions on*, 24(12):5369–5378, 2015. 2

[47] Vinay P. Namboodiri, Subhasis Chaudhuri, and Sunil Hadap. Regularized depth from defocus. pages 1520–1523. IEEE, 2008. 3

[48] Carl Olsson, Johannes Ulén, and Yuri Boykov. In Defense of 3D-Label Stereo. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1730–1737, 2013. 5, 7

[49] Vladimir Paramonov, Ivan Panchenko, Victor Bucha, Andrey Drogolyub, and Sergey Zagoruyko. Depth Camera Based on Color-Coded Aperture. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 910–918, 2016. 3

[50] Alex Paul Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):523–531, 1987. 3

[51] Nico Persch, Christopher Schroers, Simon Setzer, and Joachim Weickert. Physically inspired depth-from-defocus. *Image and Vision Computing*, 57:114–129, 2017. 3

[52] 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), 2004. 3

[53] Yoav Schechner and Nahum Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162, 2000. 4

[54] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 501–518, Cham, 2016. Springer International Publishing. 1, 2, 7

[55] Gwangmo Song and Kyoung Mu 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. 3

[56] 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. 5, 6

[57] Yuichi Takeda, Shinsaku Hiura, and Kosuke Sato. Fusing Depth from Defocus and Stereo with Coded Apertures.

In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 209–216, 2013. 3

[58] Huixuan Tang, Scott Cohen, Brian Price, Stephen Schiller, and Kiriakos N. Kutulakos. Depth from Defocus in the Wild. pages 4773–4781. IEEE, 2017. 3

[59] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *2013 IEEE International Conference on Computer Vision*, pages 673–680. IEEE, 2013. 3

[60] Michael W Tao, Pratul P Srinivasan, Sunil Hadap, Jitendra Malik, and Ravi 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. 3

[61] E Tola, V Lepetit, and P Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815,830, 2010-05. 2

[62] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 2

[63] G Vogiatzis, C Hernandez, P.H.S Torr, and R Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2241–2246, 2007. 2

[64] Ting-Chun Wang, Manohar Srikanth, and Ravi Ramamoorthi. Depth from semi-calibrated stereo and defocus. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-, pages 3717–3726. IEEE, 2016. 3

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

[66] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. pages 969–976. IEEE Publishing, 2011. 3

[67] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Computer Vision – ECCV 2018*, pages 785–801, Cham, 2018. Springer International Publishing. 2, 3

[68] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5520–5529, 2019. 3

[69] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings*, volume 07-12-, pages 4353,4361, 2015-06-01. 2

[70] Zhaokun Zhu, Christos Stamatopoulos, and Clive S. Fraser. Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109(C):47–61, 2015. 2