International Journal of Computer Vision (2022) 130:2858–2884
https://doi.org/10.1007/s11263-022-01658-w

Finite Aperture Stereo

Matthew Bailey[1] · **Adrian Hilton**[1] · **Jean-Yves Guillemaut**[1]

Received: 3 March 2022 / Accepted: 16 July 2022 / Published online: 13 September 2022

Abstract
Multi-view stereo remains a popular choice when recovering 3D geometry, despite performance varying dramatically according to the scene content. Moreover, typical pinhole camera assumptions fail in the presence of shallow depth of field inherent to macro-scale scenes; limiting application to larger scenes with diffuse reflectance. However, the presence of defocus blur can itself be considered a useful reconstruction cue, particularly in the presence of view-dependent materials. With this in mind, we explore the complimentary nature of stereo and defocus cues in the context of multi-view 3D reconstruction; and propose a complete pipeline for scene modelling from a finite aperature camera that encompasses image formation, camera calibration and reconstruction stages. As part of our evaluation, an ablation study reveals how each cue contributes to the higher performance observed over a range of complex materials and geometries. Though of lesser concern with large apertures, the effects of image noise are also considered. By introducing pre-trained deep feature extraction into our cost function, we show a step improvement over per-pixel comparisons; as well as verify the cross-domain applicability of networks using largely in-focus training data applied to defocused images. Finally, we compare to a number of modern multi-view stereo methods, and demonstrate how the use of both cues leads to a significant increase in performance across several synthetic and real datasets.

Keywords 3D reconstruction · Depth from defocus · Multi-view stereo

1 Introduction

The extraction of 3D geometry from digital images has a long and varied history, and continues to be an actively studied field. The understanding of a scene's composition is central to many applications such as robotics, augmented reality and self-driving cars; and remains a challenging problem at the forefront of computer vision research.

Given a single 2D projection of a scene captured by a camera, the problem at hand is to estimate the underlying scene structure. Though the shape of this geometry is independent to its reflectance properities prior to capture, the projection of the scene to an image entangles these quantities as pixel intensities. Clearly, this process is not invertible, and any number of solutions exist that could describe the surface manifold.

Only recently has single-image 3D reconstruction achieved compelling results due to the adoption of learning-based methods. However, such methods are not perfect, and do not necessarily generalise well to unseen data. The reasoning behind unsatisfactory results in these scenarios are not always obvious, since the perception of the scene via contextual cues, image features and shape priors are learnt as implicit components of the framework. It is therefore difficult to evaluate why one scene may work better than another, or to fully understand the limitations of a particular method. This is true of any learning-based approach to varying extents.

It is apparent then that more information is required for generalised 3D reconstruction. For passive methods, this is usually some measurable change in the scene appearance arising from different viewpoints, camera settings, or environmental conditions such as lighting. By relating these observational cues to scene-centric or camera-centric models, 3D information can be inferred from multiple 2D images.

Communicated by Matteo Poggi.

✉ Matthew Bailey
m.j.bailey@surrey.ac.uk

Adrian Hilton
a.hilton@surrey.ac.uk

Jean-Yves Guillemaut
j.guillemaut@surrey.ac.uk

[1] Centre for Vision, Speech and Signal Processing, University of Surrey, Stag Hill, Guildford, Surrey GU2 7XH, UK

For example, multi-view stereo (MVS) aims to triangulate scene points from two or more images. Within the constraints of epipolar geometry, there exists a relationship between a given scene point and its projection across multiple views; providing candidate solutions of the original 3D coordinates. The search of this candidate space is sometimes referred to as the correspondence problem, and it is solved by comparing the similarity of pixels across neighbouring viewpoints.

Clearly, this principle strongly implies a number of properties about the scene content, and subsequently the performance of such methods is intrinsically linked to the data it is operating on. While capable of sub-millimeter accuracy in the presence of uniquely textured diffused materials, MVS reconstructions begin to degrade when applied beyond this ideal Lambertian surface model. Without the introduction of regularisation or scene priors, surfaces exhibiting complex light interactions such as specularities and sub-surface scattering cannot be recovered via multi-view consistency. Furthermore, regions of low or periodic texture make correspondence challenging, with concavities and thin structures often failing due to few observations.

While many previous works focus on developing approaches that work around these constraints, in this work we aim to overcome them by generalising image formation away from the traditional pinhole camera model. In doing so, we are able to exploit the characteristics of the camera to leverage additional information about the scene.

Specifically, the use of a lens introduces aberrations to the image that are not captured by the pinhole model. Traditionally, their appearance in the image serves only as sources of outlier that must be avoided or corrected. For example, while lens distortion can be detected during calibration, defocus blurring as a result of a finite aperture cannot. While this may not be a concern for large scale scenes such as buildings, the scene must nonetheless remain within the depth of field of the camera (DoF) where pixels can be considered sharp.

At first glance, it appears the formation of defocus is the prohibitive factor here; whose corruption of the scene radiance prevents the application of MVS. While this is true it is simply unavoidable - yet strangely presents an advantageous situation. Perhaps surprisingly, defocus itself can be considered a rich source of information about the scene structure. At its simplest, this could be the location of the focal plane according to a focus measure, which identifies focused pixels according to the axiom; defocus blur acts as a low-pass filter.

Instead, in this work we pursue the analysis of the defocus appearance, which in literature is best known as depth from defocus (DFD). While MVS introduces information through changes in viewpoint, DFD instead modifies the camera parameters; such as the focusing distance or aperture size. Defocus analysis is therefore monocular, and permits the recovery of view-dependent materials that would otherwise be challenging for MVS. However, for this reason traditional implementations only achieve partial reconstructions.

In this paper, we explore how MVS and DFD can be used together to recover geometry from macro-scale scenes with complex materials; and how the combination of these cues achieves higher quality reconstructions than if they were used individually. Though some previous works have demonstrated this, no work that we are aware of does so in the context of general 3D reconstruction. As part of our evaluation, we compare against and outperform a number of modern MVS approaches.

The majority of defocus-based literature hinges on the thin lens camera model. Here, we explain why this model is not robust enough for multi-view reconstruction, and instead develop our framework around the principles of a thick lens. To supplement this, we propose a novel and practical thick lens calibration procedure suitable for macro-lenses. The effectiveness of our calibration is demonstrated experimentally on a number of real-world datasets.

Defocus-based literature has only recently started shifting towards modern learning-based methodologies. Here, we evaluate the advantages of a feature-based cost function derived from a pre-trained convolutional neural network (CNN). Though networks trained end-to-end may have ambiguity as a whole regarding generalising to different inputs, it has been shown that image-based feature extraction transfers well across domains. Our results concur with these findings, and demonstrate a step improvement over traditional pixel-based comparison. Significantly, our results indicate the feature extractor pre-trained largely on pinhole images has the same capability with defocused images.

To summarise, this paper revisits the key aspects of image-based geometry recovery - image formation, calibration and multi-view reconstruction, and presents the following contributions:

1. An MRF-based reconstruction framework unifying stereo and defocus cues using deep features
2. A novel thick lens calibration procedure used to capture a number of real-world multi-view, multi-focus datasets
3. An extensive evaluation demonstrating the benefits of our approach, including an ablation study and comparisons against several modern MVS methods
4. Real and synthetic datasets released with this paper

This paper builds on our previous works Bailey and Guillemaut (2020); Bailey et al. (2021), and combines them to produce a complete pipeline for recovering geometry from finite aperture images using stereo and defocus cues. We introduce the feature-based cost function and have included an extended calibration derivation and evaluation to better

illustrate the contributions of each cue under different conditions.

The remainder of this paper is structured as follows. Section 2 discusses previous work. Section 3 discusses image formation, and introduces the thick lens model. Section 4 explains our proposed calibration for this model, and Sect. 5 provides details on the reconstruction pipeline. Section 6 evaluates results on both synthetic and real data, and Sect. 7 concludes this work.

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 combination. 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. Moeller et al. (2015).

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 Vogiatzis et al. (2007); Logothetis et al. (2019); Hornung and Kobbelt (2006); Kar et al. (2017); Choy et al. (2016) and mesh-based Li et al. (2016); Delaunoy and Pollefeys (2014) representations allow for a globally optimal result, since all views can be evaluated jointly. Alternatively, view-dependent methods Schönberger et al. (2016); Tola et al. (2012) 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) Li and Zucker (2010) and normalised cross correlation (NCC) Li et al. (2016); Bradley et al. (2008); Furukawa and Ponce (2010). Some works exploit perspective distortion to also estimate surface normals Bradley et al. (2008). More recent approaches generally use feature descriptors to extract richer information from the source images. Though initially handcrafted Tola et al. (2010, 2012), the advent of deep learning introduced data-driven feature extraction with CNNs Zagoruyko and Komodakis (2015); Yao et al. (2018).

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 Zhu et al. (2015); Liu et al. (2010); Tola et al. (2012). Notably, PMVS Furukawa and Ponce (2010) combine matched patches rather than point clouds, and refine the final mesh using an energy optimisation to impose smoothness constraints. COLMAP Schönberger et al. (2016), 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 Ji et al. (2017) introduced the first method trained end-to-end based around a voxel grid. DeepMVS Huang et al. (2018) instead generates a plane sweep volume and aggregates matched features from an arbitary number of images. MVSNet Yao et al. (2018) introduces differentiable homography warping, and R-MVSNet Yao et al. (2019) improves the memory efficency with a recurrent architecture. PointMVSNet Chen et al. (2019) adopts a coarse-to-fine approach with multi-scale features. CasMVSNet Gu et al. (2019) develops a memory efficent cost volume and adapts it to existing methods. VisMVSNet Zhang et al. (2020) considers per-pixel visibility according to pair-wise observations and generates a cost volume via uncertainty maps. Though not advertised as MVS, neural radiance fields Mildenhall et al. (2020) achieve dense implicit reconstructions. Other notable works include Luo et al. (2019); Kuhn et al. (2020).

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 Chakrabarti and Zickler (2012); Anwar et al. (2021); Carvalho et al. (2019); Kashi-

wagi et al. (2019), 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 Tao et al. (2013). 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 Pentland (1987); Martinello et al. (2015); Song and Lee (2018) or focusing distance Favaro et al. (2008); Namboodiri et al. (2008). Changing the aperture size is often simpler, but the scene reconstruction volume is limited due to the relative blur exhibiting a symmetrical transfer function Mannan and Langer (2015). Although focal stacks largely overcome this ambiguity, refocusing the camera in this way introduces scale and translational differences between images and subsequently requires correcting Watanabe and Nayar (1998); Tang et al. (2017); Ben-Ari (2014); Bailey and Guillemaut (2020). Some methods Hasinoff and Kutulakos (2009) 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 Watanabe and Nayar (1998); Favaro (2010) and Gaussian Favaro et al. (2008); Ben-Ari (2014); Persch et al. (2017) functions. These methods do not consider many of the aberrations present in optical systems, so some works Kashiwagi et al. (2019); Martinello et al. (2015) instead directly measure the blurring response of the camera. Other works do not model the PSF explicitly, instead depending on a data driven approach Hasinoff and Kutulakos (2009); Carvalho et al. (2019); Favaro and Soatto (2005). In many cases, a thin lens defocus model is assumed despite the fact this model does not hold in real-world optical systems. Lin et al. (2013) improves reconstruction accuracy through iterative refinement. Paramonov et al. (2016) considers a model beyond a thin lens, and formulates sub-aperture disparity relative to the entrance pupil in a colour coded-aperture camera. Bailey and Guillemaut (2020) proposes a formal calibration of a thick lens camera model, and applies it to capturing and reconstructing multi-view focal stacks.

Aside from Emerson and Christopher (2019) 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, Bailey and Guillemaut (2020) 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 Li et al. (2010); Rajagopalan et al. (2004); Chen et al. (2015); often expressing the relative blurring kernel in terms of pixel disparity. Takeda et al. (2013) applies coded apertures in this way. Acharyya et al. (2016) instead uses defocus to constrain stereo matching. Other methods apply single-image defocus constraints to better recover discontinuities Wang et al. (2016); Gheţa et al. (2007).

Alternative to pairwise-stereo, some methods use lightfield cameras to combine cues Lin et al. (2015); Tao et al. (2013); Tao et al. (2017), though reconstructions are limited to a very narrow baseline. Bhavsar and Rajagopalan (2012) considers multiple viewpoints, but does not apply this to 3D reconstruction. Chen et al. (2017) 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 Chen Li et al. (2016), stereo Wu et al. (2011) and both Tao et al. (2017) 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, and therefore do not recover a complete representation of the scene. In comparison to our previous works; though Bailey and Guillemaut (2020) remains the only method we know of that achieves 3D reconstruction using only defocus cues, it forgoes the explicit multi-view consistency of MVS. Bailey et al. (2021) demonstrates the advantages of using both stereo and defocus cues in 3D, but does not use a robust cost function or extensively illustrate the contribution of each cue. In this paper, we present the complete pipeline of our thick lens-based reconstruction approach, and address these shortcomings.

3 Image Formation

3.1 Projection Model

As discussed in the introduction, the pinhole camera model has become a key component to the theory behind MVS. Ignoring lens distortion, the projection of a world-space coordinate $\mathbf{X}$ to an image point on the camera sensor $\mathbf{x}$ is described

by this model as Hartley (2000),

$$\mathbf{x} = K[R \mid \mathbf{t}]\,\mathbf{X}. \tag{1}$$

Here, the intrinsic matrix K describes the projection itself, while rotation matrix R and translation vector $\mathbf{t}$ define the camera orientation in space. We define K as,

$$K = \begin{bmatrix} F & s & x_0 \\ 0 & F & y_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

with F denoting the effective focal length, x_0 and y_0 describing the principal point (the centre of the image sensor relative to the centre of projection), and s the skew factor which is usually zero.

Let $r(\mathbf{x})$ define the radiance of the projected point. For a pinhole image, it is enough to simply assign the pixel colour according to r, as is the assumption in MVS. However, a more general expression can be used instead to describe the formation of a pixel $\mathbf{y}$ on image I Favaro et al. (2008),

$$I(\mathbf{y}) = \int k(\mathbf{y}, \mathbf{x})\, r(\mathbf{x})\, d\mathbf{x}. \tag{3}$$

Here, $k(\mathbf{y}, \mathbf{x})$ represents the PSF, or the influence of the lens with respect to the formation of defocus. The pinhole model therefore becomes a special case of Eq. 3 where k assumes a Dirac delta centred around $\mathbf{y}$; thereby permitting the captured image to represent the incident radiance.

3.2 Defocus Model

In most DFD approaches, Eq. 3 is approximated as a convolution which imposes a fronto-parallel assumption about the scene. Although this technically becomes invalid at discontinuities, we found in our experiments the error introduced is negligble. Let us define k_σ as a convolutional blurring kernel that estimates the PSF of the camera. Our image formation model then becomes a spatially variant convolution between the projected radiance and k_σ Favaro et al. (2008); Favaro (2007)

$$I(\mathbf{y}) = (k_\sigma * r)(\mathbf{y}). \tag{4}$$

The PSF kernel k_σ resembles the shape of the aperture, and describes the distribution of light formed on the sensor in defocused regions. A popular choice in literature is to approximate the PSF as a 2D Gaussian function Favaro et al. (2008); Ben-Ari (2014). In this work, we also adopt this approach and define

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

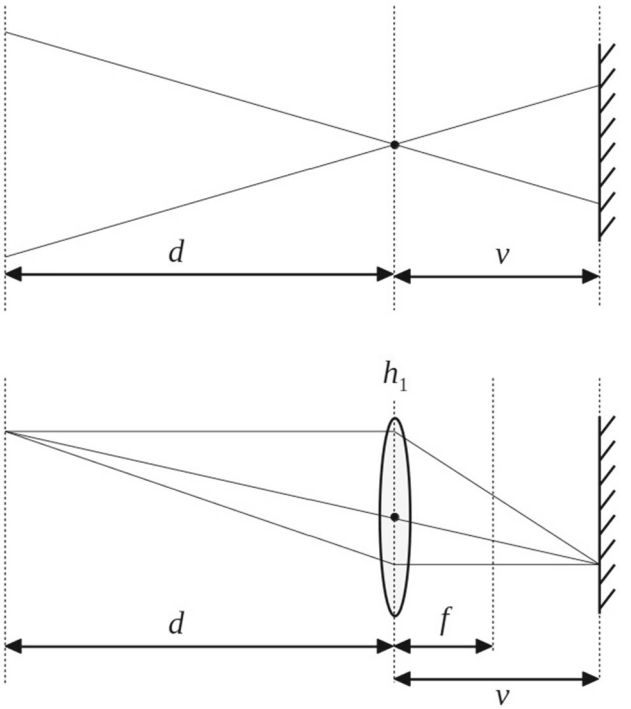

Fig. 1 Comparison of pinhole (top) and thin lens (bottom) image formation models. Thin lens assumptions introduce defocus aberration by replacing the virtual pinhole with a principal plane h_1 at the same location. In both cases, the image distance v becomes the effective focal length derived in traditional stereo camera calibration

To complete our defocus camera model, we now need to derive the blur variance σ. This aspect of the defocus model is arguably the most important, since it relates the blurred appearance to scene depth d. Most existing literature consider a thin lens abstraction of the camera optics giving Favaro (2007)

$$\sigma(d) = \frac{\gamma a v}{2}\left(\frac{1}{d} + \frac{1}{v} - \frac{1}{f}\right), \tag{6}$$

where f is focal length, a is the aperture radius, v is the sensor distance from the lens, and γ is a camera-specific constant.

This model has a number of drawbacks. First, the lens is assumed to be infinitesimally thin - simplifying the light transport to refract only once as it passes through the camera optics. In reality, light refracts at the boundary between two materials with differing refractive indices. For any physical glass lens suspended in air, light refracts once when it enters and again when it leaves.

Second, the thin lens model makes implicit and incorrect assumptions about the location of the principal plane of refraction. Comparing to the pinhole camera model, thin lens theory implies the centre of projection aligns with this refractive plane as seen in Fig. 1. This does not hold in practise, especially with macro-lenses. The implications of

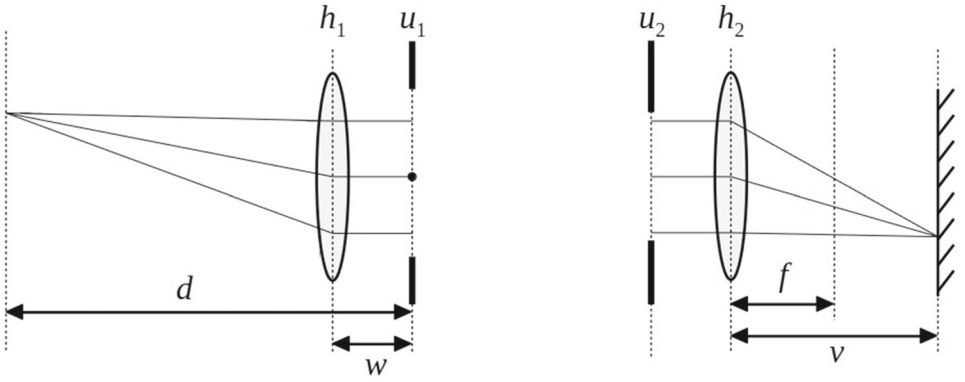

Fig. 2 Our camera model is a thick lens composed of two thin lenses each with focal length f separated by some distance. The effective pinhole location is at the entrance pupil u_1. Calculation of the defocus radius σ for a given pixel is performed relative to the principal planes h_1 and h_2

this become clear after realising defocus-based reconstructions are relative to the thin lens; while camera orientation and stereo-based reconstructions are relative to the pinhole. Therefore, any disparity between the locations of these quantities will introduce ambiguity between cues.

To overcome these problems, we model defocus formation according to thick lens principles. This model describes the camera lens as two principal planes h_1 and h_2 seperated by some distance as illustrated in Fig. 2; implying light refracts twice as it passes through the lens. Immediately, this addresses the first problem with the thin lens model.

This addition of another refractive plane in our model gives rise to a question - where is the aperture located? The answer to this is not immediately clear, but for our purposes this doesn't matter. Instead, we need only consider the virtual images of the aperture as seen through the front or the back of the camera lens. These images are referred to as the entrance u_1 and exit u_2 pupils respectively, and control the amount of light entering or leaving each lens in the model.

If the pupil diameters are the same (a symmetric lens), then their positions converge on their respective principal planes. A more realistic model considers the scenario when their sizes differ, which has the effect of displacing the pupils. The size of this displacement is proportional to the ratio of pupil diameters, or pupillary magnification p Rowlands (2017)

$$p = \frac{u_2}{u_1}. \tag{7}$$

Given that the effective pinhole location exists at the entrance pupil u_1, the second problem with the thin lens model can be addressed by finding p. Then, the displacement w of the front principal plane h_1 can be found Rowlands (2017)

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

To account for this offset, Eq. 6 now becomes

$$\sigma(d) = \frac{\gamma a v}{2}\left(\frac{1}{d-w} + \frac{1}{v} - \frac{1}{f}\right), \tag{9}$$

with scene depth d relative to the entrance pupil, and defocus observations relative to h_1. From the above it is clear that when $p \to 1$, $w \to 0$. Only under these conditions do thin lens assumptions become valid.

3.3 Camera Model Summary

Throughout this section, we have discussed three image formation models; pinhole, thin lens and thick lens. Hopefully, it is now apparent that the thick lens approach we take in this paper is a generalisation of the thin lens model; and by extension, thin lens assumptions are a generalisation of the traditional pinhole camera. In this way, it is interesting how the complexity of each model progresses by incrementally building on the principles of the previous one.

Could this dynamic be continued further, and would it be of any benefit? Consider, modern lenses are incredibly complicated pieces of equipment, with many optical elements involved in resolving the focused image. Surely, by incorporating more parameters into our model we could describe the camera behaviour with even higher precision? Certainly, additional factors could be included in the formation model - for example, we only consider light as a particle and disregard wavelength-dependent effects such as chromatic aberration. However, the majority of the complexity in modern lenses is to correct for such aberrations, so their appearance is far less significant than defocus blurring. It is not unreasonable to assume this will only continue to improve in future cameras, whereas defocus formation remains unavoidable. Moreover, the complexity of camera lenses is so great that we would argue only data-driven models, rather than our analytical model, can incorporate these less prominent features with any accuracy. That being said, we have already demonstrated in our previous work Bailey and Guillemaut (2020) the advantages our thick lens model has over traditional defocus analysis. For the scope of this paper, thick lens principles are sufficient for unifying stereo and defocus cues.

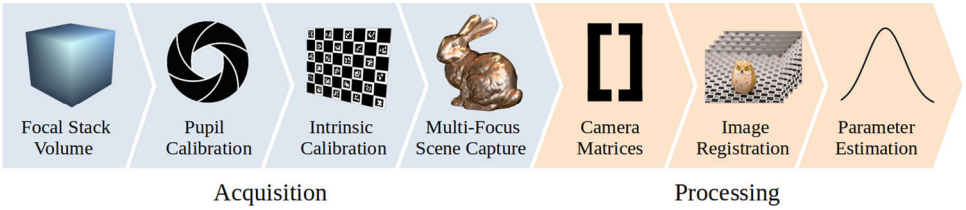

Fig. 3 The capture of a dataset is composed of two stages: aquisition and post-processing. During acquisition, an appropiate number of focal stack images are captured depending on the scene volume and aperture setting of the camera. Then, a series of images are taken concerning the thick lens calibration detailed in Sect. 4. Finally, with each camera setting calibrated, the capture of the actual scene can commence - with multi-view datasets achieved by orbiting the single camera around the scene. Using this data, the camera matrices describing image projection and pose are derived; with differences in image scale and translation as a result of refocusing the camera corrected. Finally, the parameters for the thick lens defocus model are estimated and refined

3.4 Cue Considerations

Let us now revisit Eq. 3, which describes the behaviour of all three camera models. Fundamentally, neither stereo nor defocus cues model 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. Both cues are therefore dependent on the scene appearance alone as observed in the 2D projection. This is in contrast to shape from shading methods, that aim to recover geometry with consideration of the lighting conditions; and are well known for their independence of texture. Why then do we consider two cues that appear to depend on similar information?

First, it should be re-iterated that defocus information is monocular, and therefore remains coherent in the presence of view-dependent materials. On the other hand, MVS relies on multi-view consistency, and therefore degrades in performance when applied to materials exhibiting complex reflectance. Unlike shading information defocus is a camera-centric phenomena, and its reconstruction principles can be generalised across many complex environments and scenes with little regard to their content. Often, shading cues must make assumptions about the environment such as the number of light sources and may impose restrictions on the scene materials. Provided sufficent defocus-variant texture is present Favaro (2007), we argue defocus is one of the richest passive sources of information regarding the scene structure. At the macro-scale magnification explored in this paper, this texture limitation is not a concern.

4 Calibration

The calibration of the thick lens camera model is non-trivial for several reasons. First, unlike most approaches, we do not consider camera parameters provided by the manufacturer to be accurate for all focus settings. Rather, we only consider these values relevant when the camera is focused at infinity. Secondly, to our knowledge there is no standard approach for reliably calculating the pupil ratio p, whose value is of significant importance in our model. Finally, our calibration needs to correct for translation and scale differences between multi-focus images without dependance on DoF or texture content.

In this section, we will discuss how we solved these problems. We begin by defining a number of focus settings that sweep through the scene volume. In general, the more focus settings captured, the better our model can be applied to defocus-based reconstruction. Our calibration approach can then be broken down into several stages as summarised in Fig. 3. For each setting, the following key steps are made:

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

From here onwards, we refer to parameters related to the i^{th} focus setting of this focal stack with a subscript. Without loss of generality, let us define a reference setting at $i = 0$.

4.1 Camera Matrices

In this first step, we derive the intrinsic calibration of the camera using a standard approach proposed in Zhang (2000). A calibration plane is positioned in multiple orientations and captured for each focus setting. Images are taken with both a small and a large aperture. For each setting, feature points $\mathbf{c}$ are identified from the smaller aperture images. The intrinsic matrix K_i and lens distortion coefficents for each setting are solved by minimising the reprojection error. In the following sections, images have lens distortion removed. R and $\mathbf{t}$ are calculated in a similar way for each viewpoint, using a set of scene features common to all views.

4.2 Registration

This step aims to register all images in a focal stack to a reference setting. A naive approach may be to directly use the parameters from the geometric calibration. Since F_i is related to the projection magnification m_i by Rowlands (2017)

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

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

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

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

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

where $\mathbf{c}_0$ and $\mathbf{c}_i$ are the feature coordinates, and $\bar{\mathbf{t}}_i$ is the mean of $\mathbf{t}_i^k \; \forall \, k$. Eq. 11 is solved as a function of s_i using gradient descent. Once s_i has been optimised, the corresponding $\bar{t}_i$ represents the required 2D translation. Images in the focal stack are then subject to the affine transform

$$T_i = \begin{bmatrix} s_i & 0 & \bar{t}_{ix} \\ 0 & s_i & \bar{t}_{iy} \end{bmatrix}. \tag{13}$$

After registration, all images in the focal stack share the camera matrices of the reference setting.

4.3 Parameter Estimation

In this section, we discuss how the parameters in Eq. 9 f_i, a_i, v_i and w are estimated. All parameters are implicitly assumed to be positive. We begin by calculating two intermediate variables m_i and p_i.

Pupillary Magnification: Consider images of a uniform plane focused at infinity and at each of the defined focus settings (see Fig. 4). Our approach relates the change in observed brightness in these images to the pupil ratio p_i at a particular focus setting under the following conditions:

Assumption 1 Exposure time and global illumination remain constant between the images.

Fig. 4 Calibration images of a uniform plane used for deriving average brightness focused at infinity (left) and at a focus setting (right). Besides the focus distance, all camera parameters and lighting conditions remain constant in both images. The observed change in brightness is therefore attributed to the pupil ratio. Images are white balanced and brightened for visualisation

Assumption 2 The pupil ratio has a value of 1 only when focused at infinity.

The amount of light incident to the image plane of the camera is related to the area of the smallest pupil. Therefore, assumption 2 implies that the maximum brightness observed will be when the camera is focused at infinity, since neither pupil is constricting the light entering the lens. Therefore, the value of p_i will either be greater than or less than 1 depending on the camera lens. We will assume this is unknown, and show the derivation for $p_i < 1$ where $u_2 < u_1$. From assumption 1, the following must hold true:

$$\frac{b_\infty}{b_i} = \left(\frac{u_{2\infty}}{u_{2i}}\right)^2. \tag{14}$$

Here, b_∞ and $u_{2\infty}$ are the average brightness and exit pupil diameter focused at infinity; and b_i and u_{2i} are the average brightness and exit pupil diameter at a given focus setting. Since $u_{1\infty} = u_{2\infty}$, Eq. 14 can be rewritten in terms of the entrance pupil according to Eq. 7

$$\frac{b_\infty}{b_i} = \left(\frac{u_{1\infty}}{u_{1i} \; p_i}\right)^2. \tag{15}$$

Knowing that Rowlands (2017)

$$u_1 = \frac{F_i}{N_i}, \tag{16}$$

$$u_{1\infty} = \frac{f_\infty}{N_\infty}, \tag{17}$$

where f_∞ is the known focal length when focused at infinity, N_∞ is the reported f-stop of the aperture and N_i is the effective f-stop setting; Eq. 15 can be rewritten as:

$$\frac{b_\infty}{b_i} = \left(\frac{f_\infty N_i}{F_i N_\infty p_i}\right)^2. \tag{18}$$

Since Rowlands (2017)

$$N_i = N_\infty \left(1 + \frac{m_i}{p_i}\right), \tag{19}$$

Equation 18 can be rearranged as a quadratic function of p_i by substituting Eq. 19:

$$\frac{F_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}\,p_i^2 - p_i - m_i = 0. \tag{20}$$

The value of p_i when $u_2 < u_1$ is given by the roots of Eq. 20.

$$p_i = \frac{f_\infty}{2F_i\sqrt{\frac{b_\infty}{b_i}}}\left(1 \pm \sqrt{1 + \frac{4F_i m_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}}\right). \tag{21}$$

By definition, $b_\infty > b_i$ and $F_i > f_\infty$. As a result, the discriminant of Eq. 21 will always be greater than 1 which would render a negative solution. This is therefore discarded, leaving the single positive solution of p_i,

$$p_i = \frac{f_\infty}{2F_i\sqrt{\frac{b_\infty}{b_i}}}\left(1 + \sqrt{1 + \frac{4F_i m_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}}\right). \tag{22}$$

Note here that Eq. 22 is only defined for $p_i < 1$. A similar derivation can be made for $u_2 > u_1$ by removing p_i from Eq. 15. Conversely, in this case $p_i \geq 1$:

$$p_i = \frac{m_i}{\frac{F_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}} - 1}. \tag{23}$$

Equations 22 and 23 represent a piecewise function describing the pupil ratio. The choice of either one when calculating p_i is simply a case of whichever one gives a valid solution. See Appendix 1 for a proof that only one of these solutions is always valid. The only unknown here is m_i, which we derive next.

Projection Magnification The magnification m_i in this context is the ratio of the object size in the scene to the projection of that object on the camera sensor. For a given focus setting, this is found by first finding the focusing distance d_i. This is the distance from the camera pinhole to the centre of the DoF. m_i and d_i are related as follows Kingslake (1992)

$$m_i = \frac{F_i}{d_i}. \tag{24}$$

To calculate d_i, we apply the Sum Modified Laplacian (SML) Nayar and Nakagawa (1994) focus measure to the large aperture calibration pattern images captured in Sect. 4.1. Since the poses of the patterns are known, feature points on the calibration plane can be sampled and the distance to the camera found. Regions where a high response is measured indicates an area in-focus. Assuming the DoF is a parallel plane, samples from multiple calibration images can be collected to improve robustness. The weighted mean of the distribution above a threshold gives the value of d_i, from which m_i is found.

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

$$f_i = \frac{F_i}{\left(1 + \frac{m_i}{p_i}\right)}. \tag{25}$$

Aperture The aperture radius a_i is given by Kingslake (1992)

$$a_i = \frac{F_i}{2N_i}. \tag{26}$$

Image Distance Usually, v_i is defined by Kingslake (1992)

$$v_i = f_i(1 + m_i). \tag{27}$$

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

$$v_i = f_i(1 + m_i) - (f_0 - f_i) = f_i(2 + m_i) - f_0. \tag{28}$$

Equation 28 offsets Eq. 27 by the difference in focal length relative to f_0. Essentially, this adjustment aims to ensure the principal planes of each setting align with one another.

Pupil Displacement Finally, we can now define the value of w according to Eq. 8.

4.4 Parameter Refinement

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

We optimise w using scene features with known position in the world reference frame. Our cost function is based on

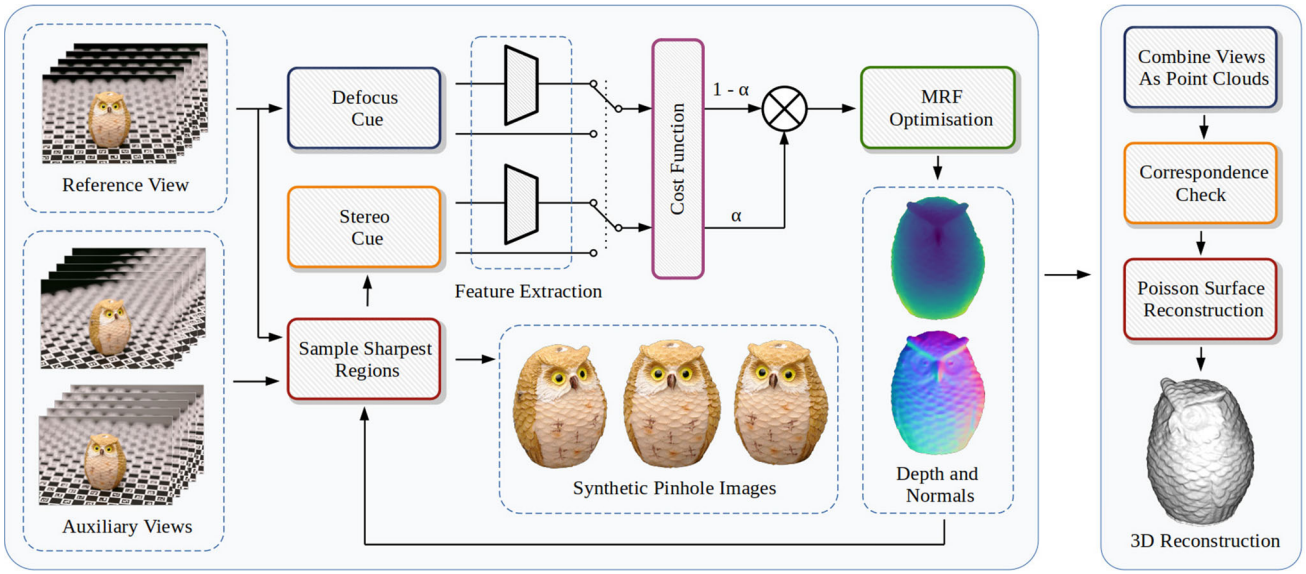

Fig. 5 Diagram of the proposed iterative reconstruction framework. Defocus and stereo observations are generated from the calibrated focal stacks and synthetic pinhole images respectively. The cost function is generated from these observations via pixel values or extracted features, and weighted according 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 fused together

the relative blur between pairs of images in the focal stack. The cost function presented here is similar to the one used in Sect. 5 for reconstruction. First, we define the relative blur between settings i and j:

$$\sigma_{ij}(d) = \sqrt{|\sigma_i(d)^2 - \sigma_j(d)^2|} \tag{29}$$

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

$$\min_{w} \sum_{\{ij\}\in\Omega} \sum_{k} \| \sigma_{ij}(d^k) \circ I_a \; - \; I_b \|^2 \tag{30}$$

$$\{a, b\} = \begin{cases} \{i, j\} & \sigma_i(d) < \sigma_j(d) \\ \{j, i\} & \text{otherwise} \end{cases} \tag{31}$$

Here, Ω is a vector of paired image indices, $\circ$ is a defocus operator which we define later, and d^k is the distance of the k^{th} feature from the camera. Equation 30 blurs whichever image is sharpest to match the other for every feature, and compares the result with a pixel-wise square difference. This sparse optimisation can be thought of as a per-viewpoint global adjustment of all blurring functions describing the focal stack.

5 Reconstruction

Our 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, as shown in Fig. 5.

Per-View Reconstruction Using stereo and defocus cues, we reconstruct per-viewpoint depth maps. As input, we take multi-view focal stacks captured and calibrated using the approach discussed in Sect. 4. These images have a narrow DoF, making them unsuitable for direct stereo matching. As part of our pipeline, we infer a focused image according to the current depth estimate, and perform stereo matching on these synthetically generated 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. Selectively composite the focal stack inputs 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 Kazhdan and Hoppe (2013) to generate the final triangular surface mesh.

5.1 Energy Function

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,

$$E(\mathbf{x}, n) = (1 - \alpha(n)) \sum_{p \in \nu} \Phi_D(x_p) + \alpha(n) \sum_{p \in \nu} \Phi_S(x_p) + \frac{\lambda}{2^{n-1}} \sum_{(p,q) \in \epsilon} \Psi_{pq}(x_p, x_q). \tag{32}$$

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. The proposed method linearly modulates 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 Olsson et al. (2013).

In our framework, we assume each pixel represents a surface and model it as a tangent plane. The orientation of each surface is estimated after every iteration by fitting a plane to neighbouring 3D points via singular value decomposition. 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 32 is minimised using α-expansion Boykov et al. (2001); Szeliski et al. (2008).

Each data term depends on a photometric cost function that compares the similarity of two image patches. In this paper we evaluate two such functions. For now, this is denoted Δ and will be explained in further detail later. We will now define each of the terms in Eq. 32.

5.2 Defocus Term

To calculate the defocus term for a pair of images $\{I_i, I_j\}$ in a given focal stack, a scale-space approach is taken. The relative blur between the images is found according to Eq. 29, and the sharper image is blurred to match the other. The cost function $\phi_D(x_p)$ is defined by the similarity between the defocused and original image

$$\phi_D(x_p) = \sum_{\{ij\} \in \Omega} \Delta\left(\sigma_{ij}(x_p) \circ I_a,\ I_b\right). \tag{33}$$

As in Eq. 30, $\circ$ denotes the defocus operator, $\mathbf{\Omega}$ contains indices of paired images, and $\{a, b\}$ are defined in Eq. 31. Since the accuracy of DFD is greatest when relative blur is small, only neighbouring images in the stack are paired together. When evaluating Eq. 33, we first remove harmonic texture components in the source images

$$I_i = I_i - (I_i \circ k_\sigma). \tag{34}$$

This procedure, proposed in Favaro (2007), removes defocus-invariant texture components, and has been shown to improve the performance of focus analysis. We define our defocus operator $\circ$ as a linear diffusion operator as proposed in Favaro et al. (2008). Although this is equivalent to the Gaussian PSF discussed in Sect. 3.2, we found linear diffusion performs better with subpixel defocus radii. The forward diffusion constraint is enforced by starting Eq. 33 at the label closest to the depth d_0 where the relative blur $\sigma_{ij}(d_0) = 0$. We derive this from Eq. 29:

$$d_0 = \frac{a_i v_i \pm a_j v_j}{\frac{a_i}{f_i}(v_i - f_i) \pm \frac{a_j}{f_j}(v_j - f_j)} + w. \tag{35}$$

The above simplifies to the result in Favaro et al. (2008) when $f_i = f_j$, $a_i = a_j$ and $w = 0$. Finally, the generated cost volume is normalised according to

$$\Phi_D(x_p) = 1 - \exp\left(-\frac{\phi_D(x_p)}{\mu_D}\right), \tag{36}$$

where μ_D is the mean of the cost volume ϕ_D.

5.3 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 generated images that approximate the scene radiance.

Compared to our previous work Bailey et al. (2021), we found that deblurring the focal stacks via non-blind deconvolution tends to be a source of instabililty; primarily in regions which do not have an accurate depth estimate. Overall, the benefits of a potentially sharper image vs the unstable conseqences when the reconstruction fails were not worthwhile.

With our datasets, we found the radiance estimate produced through selective sampling of the focal stack produced a result that was perfectly adequate for stereo matching. As before, observations from either side of the reference view are used to improve robustness to occlusions.

Given an estimate of the depth map from the reference view, the surface is raymarched to determine the distance of all pixels from each view. Some pixels in the auxillary views will not intersect this surface, but this means they are probably not visible in the reference view anyway. With this estimate of the scene radiance, let us now look in detail at how a single pixel p is processed.

Assuming p is in the reference view, we define a square support patch W_p centred around p, and cast rays into world-space. Unike Bailey et al. (2021), this is not done for every pixel in the support patch - only the four corners. As a result, computational efficiency is improved dramatically, and remains reasonably consistent regardless of the patch size.

These rays are intersected with sample tangent planes corresponding to p; at 3D locations determined by the candidate labels. By considering the surface orientation in this way, perspective distortion is applied to better resemble the patch appearance in the auxillary view. Pixels are then sampled between these corners, with subpixel sampling performed via bilinear interpolation. For label x, the vector of costs defining the similarity between a patch in the reference view W_p and patches in the auxillary views $\hat{W}_p$ is defined by

$$\varphi_S(x_p) = \left\{ \Delta\left(W_p, \hat{W_p}^0\right), ..., \Delta\left(W_p, \hat{W_p}^j\right) \right\}, \tag{37}$$

where $\mathbf{\Omega_S}$ defines the vector of auxillary views with $j \in \mathbf{\Omega_S}$. In our implementation, we consider 4 neighbouring views. To improve robustness, only the best 2 scores are considered per label from $\varphi_S(x_\mathbf{p})$, and are averaged together; denoted $\phi_S(x_\mathbf{p})$. Finally, the costs are normalised to produce the final stereo term, where μ_S is the mean of the cost volume ϕ_S:

$$\Phi_S(x_p) \;=\; 1 - \exp\left(-\frac{\phi_S(x_p)}{\mu_S}\right). \tag{38}$$

5.4 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 Szeliski et al. (2008)

$$\Psi_{pq}(x_p, x_q) = \min\left(\Psi_{max}, V_{pq}(x_p, x_q)\right). \tag{39}$$

The above enforces pairwise smoothness between two pixels p and q taking labels x_p and x_q respectively, with the truncation preserving discontinuities. Following Bailey and Guillemaut (2020), we define V_{pq} as a second-order prior and exploit the tangent plane surface model. For two world-points $\mathbf{P}$ and $\mathbf{Q}$ corresponding to labels x_p and x_q respectively, we define V_{pq}

$$V_{pq}(x_p, x_q) = \left(\frac{1}{\delta(n)(N-1)}\left|\frac{(\mathbf{Q}-\mathbf{P})\cdot\mathbf{q^n}}{\mathbf{p^r}\cdot\mathbf{q^n}}\right|\right)^2, \tag{40}$$

similar to the definition proposed in Olsson et al. (2013). Here, $\mathbf{q^n}$ is the normal of the surface related to pixel q, $\mathbf{p^r}$ 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 $\lambda = 10000$.

5.5 Photometric Cost Function

Throughout this section, our photometric cost has been abstracted away as some similarity score between two image inputs. Here, let's consider the comparison of two image patches $\mathbf{a}$ and $\mathbf{b}$. In this paper, we explore two approaches in our implementation. The first takes the per-pixel sum of square differences (SSD) for all pixels contained within the patches,

$$\Delta_{SSD}(\mathbf{a}, \mathbf{b}) = \sum_i (a_i - b_i)^2. \tag{41}$$

This is a very simple and in some ways naive approach, but remains a popular choice in DFD. In fact, Eq. 41 was used exclusively in our previous work Bailey et al. (2021).

The second cost function we present in this paper is based on the extraction of learnt features. This is achieved using selected CNN layers of a pre-trained image classifier. In this work, we use ResNet-50 pre-trained on the ImageNet dataset. The model was obtained from the TorchVision package for PyTorch. It was trained using stochastic gradient descent with 10^{-4} weight decay and 0.9 momentum for 90 epochs; with a batch size of 32 and a learning rate of 0.1. Every 30 epochs, the learning rate was reduced by a factor of 0.1. All pooling layers and fully-connected layers are removed, and only the initial 7x7 convolutional layer and the first 2 bottleneck layers are used. These modifications were made because the training images are significantly larger than the image patches we wish to evaluate. Consequently, we can truncate the network and still maintain a receptive field that is appropiate for our use case.

Let the function R represent a forward pass of our ResNet-based feature extractor. Our similarity score then becomes a comparison between features instead of pixels,

$$\Delta_{CNN}(\mathbf{a}, \mathbf{b}) = \sum_i (R(\mathbf{a})_i - R(\mathbf{b})_i)^2. \tag{42}$$

Fig. 6 Materials simulated in our synthetic datasets: gold (top row), stone (middle row) and wood (bottom row)

Note that **a** and **b** remain subject to the usual image normalisation required by the network. In all experiments, we use a patch size of 11x11 pixels - an increase from the 5x5 patch size used in Bailey et al. (2021). When using Δ_{CNN}, the output of R produces a 3x3 patch with 512 channels; which is flattened to a vector containing 4608 features.

5.6 Point Cloud Fusion

To filter out significantly erroneous points in the point cloud outputs, a post-processing correspondence check is performed. This process retains corresponding points from neighbouring views that determined similar results during reconstruction, indicating a level of robustness in that region, and eliminates them otherwise. Our implementation requires each point to correspond in at least two adjacent views to within 0.5mm. We also exclude corresponding points where the difference in normal vectors exceeds 30 degrees. The position and normal vectors of all remaining points are averaged with their corresponding matches, and are subject to screened Poisson surface reconstruction to generate the final triangular mesh of the scene.

6 Evaluation

In this section we evaluate the performance of our approach on synthetic and real data. Here, we perform an ablation study to analyse the contribution of each cue, by comparing the proposed method against the performance of the stereo and defocus terms operating individually. This is achieved by fixing the weighting term α to 0 for defocus and 1 for stereo for all iterations except the first. In all experiments, the first iteration of the pipeline is defocus only to generate an initialisation of the surface and an estimate of the radiance required for stereo matching. Our evaluation considers two cost functions: sum of square differences (SSD) and a pre-trained feature-based cost (CNN). For all experiments, we process $N = 100$ labels and run for 5 iterations with a visual hull initialisation. Real-world object silhouettes were generated using Rother et al. (2004), with any ambiguous regions manually corrected. Since the focal stack images are registered during calibration, only one silhouette is necessary per view. We found that the defocus term did not respond well without the object silhouettes shrunk to remove blurring due to background pixels. Our results are therefore missing some regions around the boundaries of objects, which is particularly apparent in the Dragon dataset. Note that, in principle, our approach does not necessarily require a visual hull initialisation.

This section begins by explaining how we generated our synthetic and real datasets. Next, a per-viewpoint evaluation is performed by comparing the accuracy of the depth maps our method produces across a range of metrics. We then explore the performance of our method in a 3D context. A quantitative analysis of the synthetic data is performed on the fused point clouds where we also compare to several modern MVS methods, before a qualitative comparison on the real datasets is conducted. Finally, an ablation study is performed to analyse the effect of the number of images in the focal stack.

6.1 Datasets

6.1.1 Synthetic

To generate the synthetic data, photo-realistic images of the Stanford Armadillo, Bunny and Dragon were rendered from 24 viewpoints using Blender. Each object was rendered with 3 different materials as seen in Fig. 6; gold, stone and wood. This initial output from the renderer represents the pinhole radiance of the scene. These images then had depth of field applied by blurring them with a Gaussian PSF according to our convolutional model, with f = 100mm, a = 4.55mm and w = 0mm. Each viewpoint is processed to create a 5-image focal stack, with the focusing distance uniformly incremented according to the ground truth depth maps. To simulate image noise, Eq. 4 is modified to become

$$I(\mathbf{y}) = (k_\sigma * r)(\mathbf{y}) + \eta, \tag{43}$$

where η is modelled as additive white Gaussian noise. For these experiments, the standard deviation of η is set equal to 1% of the pixel value range. In combination with the noise-free data, this totals 18 synthetic datasets.

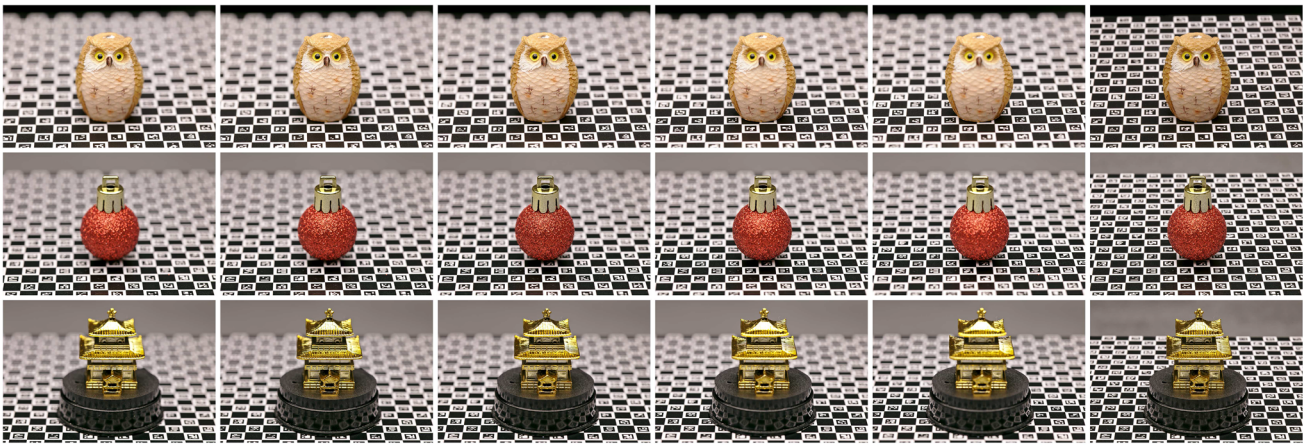

Fig. 7 Example focal stack input images for our real datasets; Owl (top row), Bauble (middle row) and Temple (bottom row), with the focusing distance increasing from left to right. Only five of nine total images for the Bauble and Temple datasets are shown. Rightmost column shows the f/22 pinhole images used by the MVS methods we compare against

Table 1 MAE of generated depth maps from datasets with 0% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	0.4677	0.5558	0.6206	0.4635	0.4893	0.5596	0.6661	0.7589	0.8146	0.5996
DFD (CNN)	**0.3428**	0.3998	0.4418	**0.3274**	0.3620	0.3862	**0.4719**	0.5017	0.5560	0.4211
Stereo (SSD)	1.7132	0.5191	0.5039	2.6894	0.3316	0.5751	2.4367	0.7152	0.6696	1.1282
Stereo (CNN)	1.2583	0.4870	0.3912	2.3017	0.2635	0.3477	2.0981	0.6856	0.5738	0.9341
Proposed (SSD)	0.7127	0.3083	0.3540	1.0115	0.2035	0.3256	0.9259	0.3666	0.4159	0.5138
Proposed (CNN)	0.4699	**0.2219**	**0.2496**	0.5590	**0.1281**	**0.1823**	0.6074	**0.2350**	**0.2801**	**0.3259**

Bold values indicate the top performing approach for each object and material
All values are given in millimetres, and represent the average error across all viewpoints

6.1.2 Real

Real-world datasets are acquired according to the procedure described in Sect. 4. In this paper we present three datasets; Owl (29 views, 5-image stacks), Bauble (18 views, 9-image stacks) and Temple (16 views, 9-image stacks). An example set of images from a single viewpoint are shown in Fig. 7. These are small objects that require relatively high magnification to photograph, and exhibit reflectance properties that resembles the synthetic data. Aperture values were chosen to be f/5.6 for Owl, and f/6.3 for Bauble and Temple. For camera pose estimation and comparison to MVS, small apertures images were taken with an f-stop of f/22.

The datasets were captured using a Canon EOS 5DS camera with a 100mm macro-lens. By physically measuring the pupil diameters as viewed from the front and back of the lens, we found the pupil ratio to be approximately 0.92 when focused at infinity - closely matching the assumptions made in Sect. 4.3. The images were downsampled to 2184 x 1464 pixels with 16-bit colour depth, before having lens distortion corrected and undergoing registration.

6.2 Depth Map Evaluation

The per-viewpoint reconstruction approach we take permits us to evaluate performance by directly analysing the generated depth maps. This allows for a more direct evaluation of the cues, since the post-processing steps required to generate a 3D model often attenuate or remove significant regions of error.

6.2.1 Synthetic

We evaluate the performance of the single-view reconstructions using Mean Absolute Error (MAE), Mean Square Error (MSE) and % Bad Pixels above 0.25mm, and take the average across all views. Tables 1, 2 and 3 show the results of this evaluation under ideal 0% noise conditions, while Tables 4, 5 and 6 show results under 1% noise. In all cases, bold indicates the top performer determined by the lowest error reported for each column; with the proposed achieving the best performance in most instances.

Table 2 MSE of generated depth maps from datasets with 0% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	1.7634	1.7150	1.9887	0.7902	0.6803	0.8507	2.1411	1.8789	2.3352	1.5715
DFD (CNN)	**1.3150**	1.1544	1.4903	**0.4956**	0.4107	0.4898	**1.5944**	1.1209	1.5067	1.0642
Stereo (SSD)	11.8569	3.9730	3.1293	18.0694	1.5366	2.0851	20.5508	6.6582	5.3865	8.1384
Stereo (CNN)	8.8662	3.7989	2.6640	18.0354	1.1104	1.2247	20.8706	7.2628	5.4120	7.6939
Proposed (SSD)	2.2276	1.5351	1.6970	2.1105	0.4671	0.6332	2.9696	1.5051	1.7846	1.6589
Proposed (CNN)	1.4470	**1.0629**	**1.3325**	0.8881	**0.2898**	**0.3604**	1.9019	**0.9489**	**1.2490**	**1.0534**

Bold values indicate the top performing approach for each object and material
All values are given in millimetres, and represent the average error across all viewpoints

Table 3 % Bad pixels from depth maps with greater than 0.25 mm error generated from datasets with 0% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	55.65	64.48	68.27	60.05	67.88	69.49	65.36	74.38	75.63	66.80
DFD (CNN)	**43.02**	57.09	58.63	**48.81**	60.60	59.14	**55.90**	67.96	67.42	57.62
Stereo (SSD)	80.24	24.07	29.99	90.49	15.63	33.27	86.89	25.88	32.01	46.50
Stereo (CNN)	68.41	21.63	19.91	81.70	12.76	19.46	78.44	22.62	23.25	38.69
Proposed (SSD)	70.70	21.19	27.63	81.52	15.77	30.39	77.88	26.57	32.88	42.73
Proposed (CNN)	55.45	**15.57**	**16.30**	61.72	**8.89**	**14.75**	63.74	**16.84**	**19.67**	**30.33**

Bold values indicate the top performing approach for each object and material
Values represent the average error across all viewpoints

Table 4 MAE of generated depth maps from datasets with 1% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	0.8830	1.7862	2.1420	1.3310	1.7808	1.9687	1.4694	2.2381	2.5746	1.7971
DFD (CNN)	0.6310	1.0146	1.6500	0.9614	0.9468	1.4938	1.0521	1.3543	2.0016	1.2339
Stereo (SSD)	1.7575	0.6204	0.5961	2.7919	0.4355	0.6527	2.5682	0.9536	0.8415	1.2464
Stereo (CNN)	1.3570	0.6891	0.7066	2.6374	0.3469	0.5643	2.5759	1.0963	0.9701	1.2159
Proposed (SSD)	0.8554	0.5685	0.6175	1.4814	0.5249	0.6836	1.2744	0.8336	0.8012	0.8489
Proposed (CNN)	**0.5576**	**0.3053**	**0.4997**	**0.9065**	**0.1991**	**0.5080**	**0.8511**	**0.4144**	**0.6371**	**0.5421**

Bold values indicate the top performing approach for each object and material
All values are given in millimetres, and represent the average error across all viewpoints

Table 5 MSE of generated depth maps from datasets with 1% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	2.8197	6.7594	9.8628	3.7143	5.7449	6.8730	5.9098	9.8261	13.0863	7.1774
DFD (CNN)	1.8713	2.6320	6.6965	2.4226	1.7354	4.8073	3.7714	4.1111	8.8884	4.1040
Stereo (SSD)	12.5764	5.4708	4.6355	19.8165	3.0136	2.9735	22.7204	10.3162	8.4367	9.9955
Stereo (CNN)	10.8243	6.5852	6.9420	24.4333	2.0313	2.9657	30.6319	13.9392	11.2732	12.1807
Proposed (SSD)	2.7353	3.0031	3.5305	4.2895	1.9686	2.1308	4.9882	4.7122	5.2661	3.6249
Proposed (CNN)	**1.6525**	**1.4068**	**2.7509**	**2.1008**	**0.5236**	**1.6627**	**2.9451**	**1.8495**	**3.3945**	**2.0318**

Bold values indicate the top performing approach for each object and material
All values are given in millimetres, and represent the average error across all viewpoints

Table 6 % Bad pixels from depth maps with greater than 0.25mm error generated from datasets with 1% noise

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
DFD (SSD)	75.98	89.52	91.54	85.15	90.56	91.97	84.69	92.13	93.61	88.35
DFD (CNN)	64.99	81.93	87.48	76.38	82.14	86.87	76.28	85.94	89.56	81.29
Stereo (SSD)	80.38	25.62	31.85	90.65	17.07	35.95	87.09	29.23	34.78	48.07
Stereo (CNN)	68.93	24.58	26.01	83.02	14.72	**27.97**	79.64	27.50	**30.82**	42.58
Proposed (SSD)	74.45	33.04	36.86	87.39	29.25	42.82	82.67	40.34	41.89	52.08
Proposed (CNN)	**60.22**	**19.48**	**25.97**	**73.14**	**12.13**	28.90	**71.30**	**22.89**	32.62	**38.52**

Bold values indicate the top performing approach for each object and material
Values represent the average error across all viewpoints

Under noisy conditions, the performance of defocus degrades significantly, yet this does not appear to negatively influence the combination of cues; the contrary in fact. At first this seems strange - noise is not explicitly modelled in either cue, so why is only defocus sensitive to it? Our understanding is as follows. The basis of defocus modelling relies on texture analysis; specifically the appearance of high frequency textures under defocused conditions. By artifically injecting additive noise to the image, defocused regions now contain a large amount of unexpected high frequencies that confuse the cost function and degrade the resulting depth map. In contrast, the stereo term does not concern itself with the spectrum of texture components; simply the similarity of two image patches. Hence, the noise only increases the variance of the cost function, and does not impact the results a great deal. Even under adverse conditions, the defocus cue appears to positively influence the proposed method; and helps us achieve the best result in almost all cases.

Figures 8, 9 and 10 provide further insight by illustrating how each cue behaves recovering depth maps in the presense of different geometry and materials. These figures show both the SSD and CNN cost functions. In ideal 0% noise conditions, defocus appears to produce complete yet imprecise reconstructions, whereas stereo achieves higher accuracy at the expense of significant outliers. In combination, a balance of these benefits is achieved. The results under noisy conditions reflect the analysis above, with defocus alone highly sensitive to noise yet the proposed continuing to perform consistently.

6.2.2 Real

Figure 11 show a selection of depth map reconstructions on the real data. As with the synthetic data, the combination of cues appears to improve the depth map consistency and reduce significant error while also extracting detailed features. This figure shows the performance of both the SSD and CNN cost functions, and while this general trend is followed by both sets of results, the CNN cost produces the smoothest and most consistent output.

6.3 3D Reconstruction Evaluation

We compare performance on our datasets to three view-dependent MVS approaches; CasMVSNet Gu et al. (2019), VisMVSNet Zhang et al. (2020) and COLMAP Schönberger et al. (2016). 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. Though all of our datasets have a 16-bit colour depth, all MVS methods require 8-bit input images instead.

To share our pinhole camera calibration with COLMAP, we manually generated the configuration files that would otherwise be created by its structure from motion pipeline. Although CasMVSNet and VisMVSNet provide scripts to convert from the COLMAP format, we found overall these methods produced the best results when configured directly with our calibration and constrained to the same auxillary views our framework uses. These methods were run pre-trained with 256 labels on an Nvidia RTX 3070 graphics card; and the input images were downsampled to a maximum resolution of 1536 x 1024. For point cloud fusion, both methods use a threshold of 3 consistent views. Otherwise, parameters were left at their default values.

In all cases, our approach uses a visual hull initialisation for the first iteration, but is then disabled for all subsequent iterations. Since the MVS methods do not have access to our silhouette information, for fairness all points that are background are removed. For the synthetic data, this is based on the RGB value of the reconstructed point cloud. The real-world scene reconstructions are instead cleaned based on the position of the points; with the majority of points outside of the object volume removed. We do not perform similar post-processing on our own results. Finally, normals for the CasMVSNet point clouds were estimated prior to Poisson surface reconstruction since CasMVSNet does not provide this.

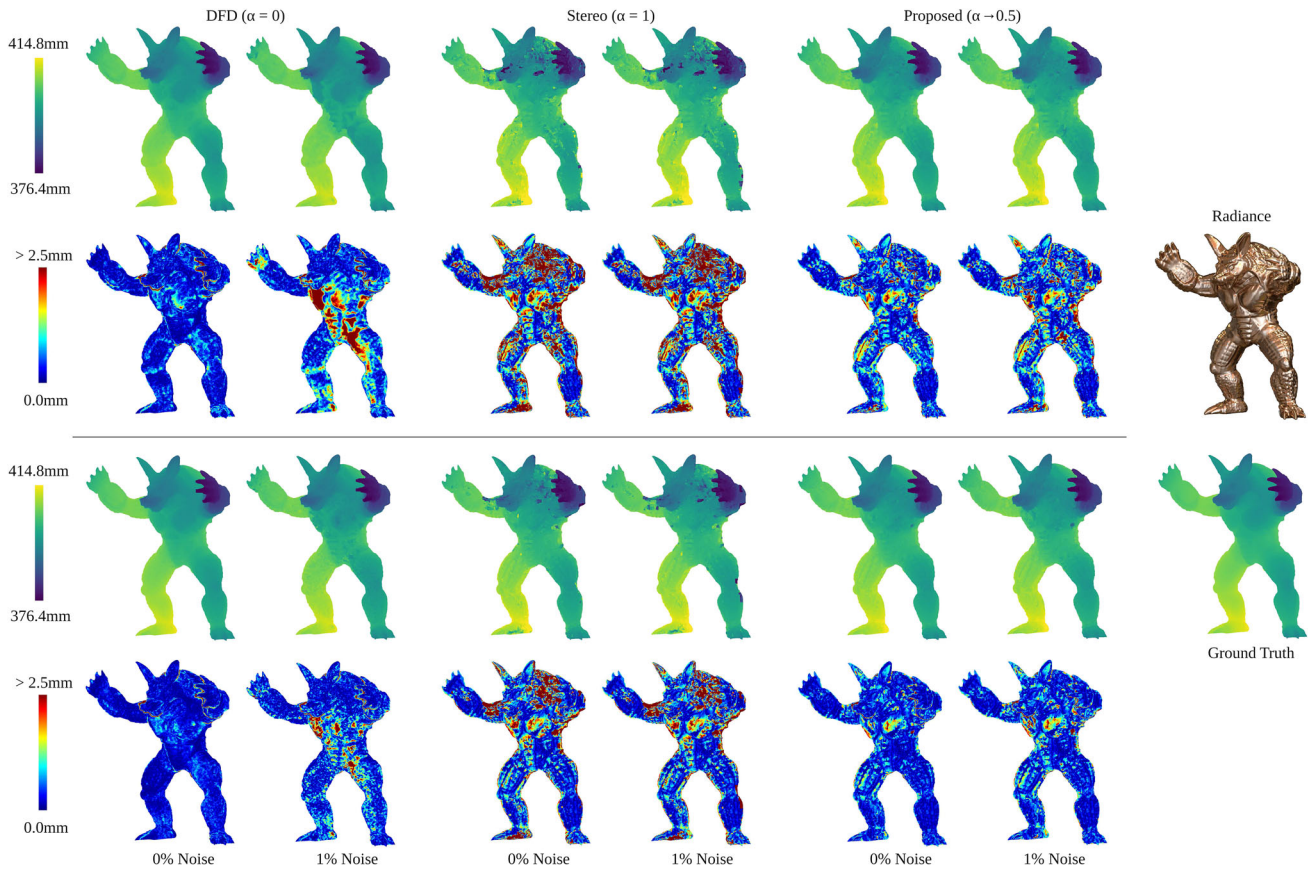

Fig. 8 Single view results on the gold Armadillo dataset with 0% and 1% noise. Rows 1 & 2 show the results from the SSD cost function; and rows 3 & 4 show results using the CNN cost function. Odd rows: depth maps produced by each variant of the method. Even rows: error maps when compared to the ground truth. Our method demonstrates robustness to noise despite the performance of both cues degrading when used seperately

Table 7 Point cloud F-scores with $\tau = 0.5$mm

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
CasMVSNet	0.8579	0.9665	0.9672	0.7915	0.9443	0.9509	0.7755	0.9004	0.9058	0.8956
VisMVSNet	0.6466	0.9151	0.8952	0.7056	0.9327	0.9208	0.6394	0.8786	0.8685	0.8225
COLMAP	0.7720	0.9734	0.9695	0.6501	0.9540	0.9508	0.6719	0.9092	0.8984	0.8610
DFD (SSD)	0.9276	0.8834	0.8691	0.8841	0.8863	0.8487	0.8254	0.7760	0.7739	0.8527
DFD (CNN)	**0.9640**	0.9529	0.9372	**0.9285**	0.9340	0.9178	**0.8738**	0.8632	0.8530	0.9138
Stereo (SSD)	0.7838	0.9765	0.9634	0.5546	0.9576	0.9147	0.6397	0.9186	0.9038	0.8459
Stereo (CNN)	0.8719	0.9774	0.9811	0.7209	**0.9579**	0.9530	0.7542	0.9230	0.9193	0.8954
Proposed (SSD)	0.8850	0.9809	0.9765	0.7723	0.9575	0.9365	0.7880	0.9194	0.9150	0.9035
Proposed (CNN)	0.9333	**0.9826**	**0.9851**	0.8779	0.9576	**0.9575**	0.8565	**0.9269**	**0.9293**	**0.9341**

Bold values indicate the top performing approach for each object and material
Results generated from data with 0% noise

6.3.1 Synthetic

Figures 12, 13 and 14 show a comparison of 3D reconstructions on a selection of synthetic datasets with 1% noise. Recall under these conditions in the previous section defocus did not perform well. The results shown appear to indicate much the same, with the proposed depending heavily on the stereo term to produce a coherent output.

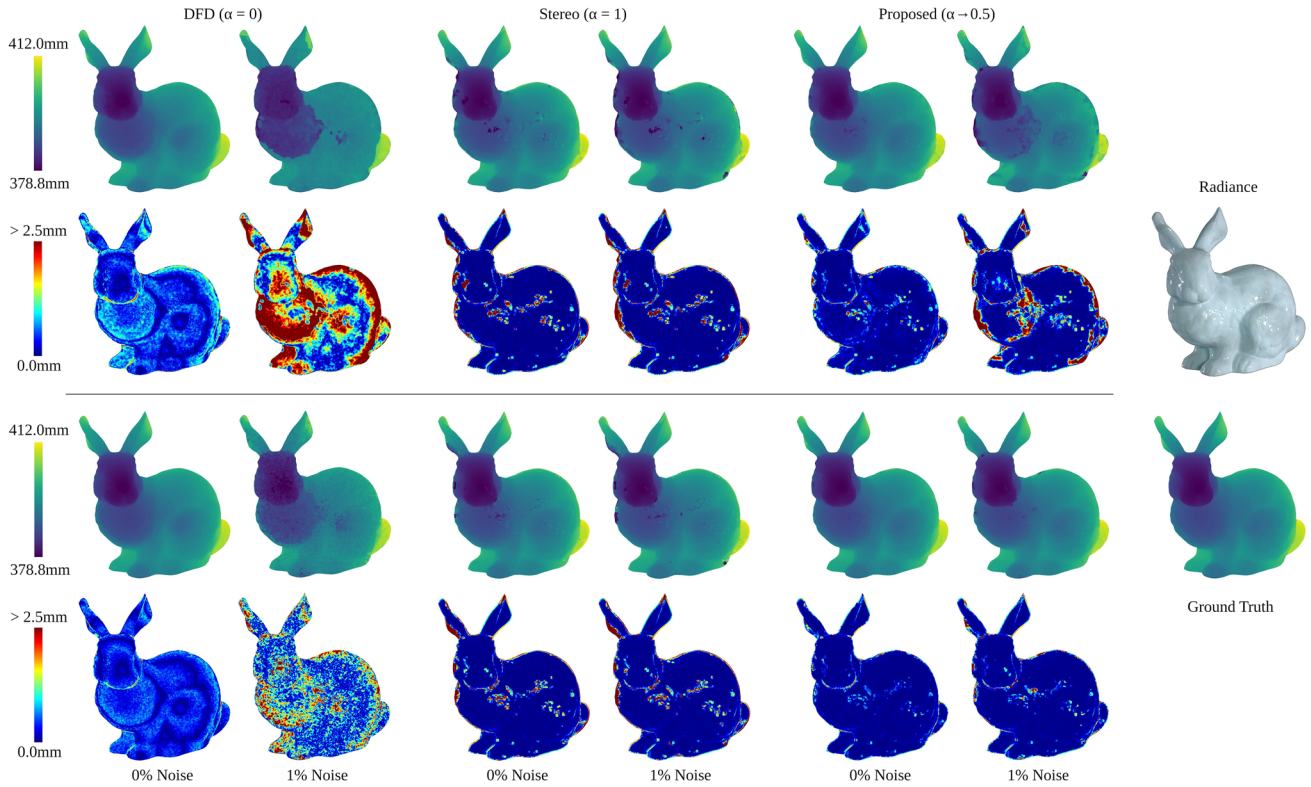

Fig. 9 Single view results on the stone Bunny dataset with 0% and 1% noise. Rows 1 & 2 show the results from the SSD cost function; and rows 3 & 4 show results using the CNN cost function. Odd rows: depth maps produced by each variant of the method. Even rows: error maps when compared to the ground truth. The proposed achieves the highest overall precision and attenuates outliers resulting from specular regions

Tables 7 and 8 show an evaluation of our synthetic data on the fused point cloud outputs using the F-score metric provided by Knapitsch et al. (2017). In both tables, bold indicates the top performer of each column which maximises the score. In ideal 0% noise conditions where both cues are functioning at their best, we outperform all MVS methods, with the proposed outperforming the individual cues the majority of the time and achieving the best result on average. Under noisy conditions, the result is less clear-cut; though the proposed remains the best performer on average. Note the consistency in performance between Table 8 and Figs. 12, 13 and 14.

Figures 15 illustrates the average recall and precision of the point clouds across all experiments with 1% noise. Figure 16 shows the same, but on the Poisson meshes. Note the proposed approach achieves the best recall of all the methods, with the individual DFD and stereo terms consistently underperforming compared to the proposed. Interestingly, the stereo term achieves greater recall than defocus term, though this is probably due to the cross-correspondence check during point cloud fusion. Observe the difference between the CNN and SSD cost functions - the former achieves better performance in all cases. While they recover less overall completeness, all MVS methods appear to outperform in terms of precision; though the difference between the proposed and MVS closes when comparing meshes instead of clouds. This is to be expected to some extent, especially when compared to the performance of our stereo term by itself. If anything, this indicates more performance remains on the table that could be exploited by improving the robustness of each cue. However, the main objective of this paper is to explore their complementary nature rather than absolute performance, so this is left as future work.

In comparison to the depth evaluation in Sect. 6.2, the complete yet imprecise nature of the defocus term is less important due to the cross consistency checks performed when generating the point cloud. However, it remains useful for recovering the geometry of complex materials. This is reflected in the F-scores, with defocus performing best on the highly specular gold material while the stereo-based methods struggle to resolve a complete cloud. The decomposition of the F-scores showed the performance of the proposed exceeds that of both terms individually in recall and precision. A similar argument could be made from Sect. 6.2 regarding the performance of the ablation under noisy conditions.

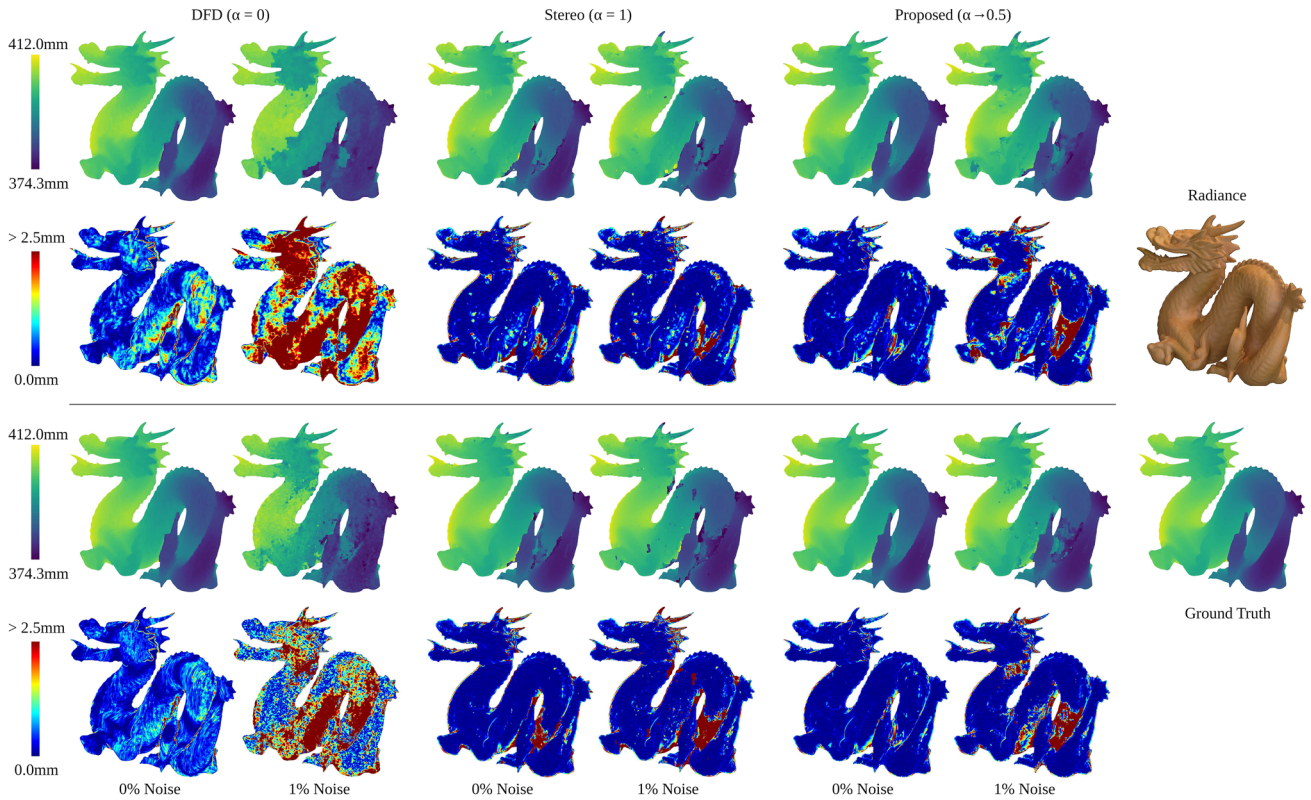

Fig. 10 Single view results on the wooden Dragon dataset with 0% and 1% noise. Rows 1 & 2 show the results from the SSD cost function; and rows 3 & 4 show results using the CNN cost function. Odd rows: depth maps produced by each variant of the method. Even rows: error maps when compared to the ground truth. Since this dataset has a largely diffused surface, failings in the stereo term are mostly due to occlusion. The proposed successfully captures the benefits of single-viewpoint reconstruction from the defocus term while retaining the higher accuracy afforded by the stereo term

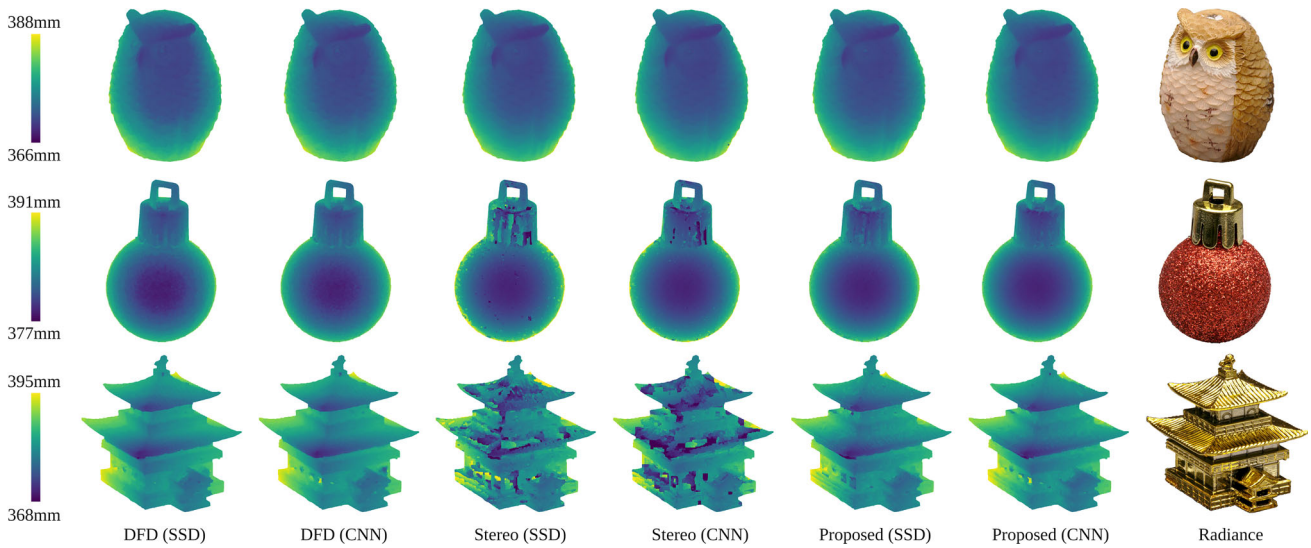

Fig. 11 Single view results on the real datasets Owl (top row), Bauble (middle row) and Temple (bottom row). Depth maps normalised manually to the specified range

Table 8 Point cloud F-scores with $\tau = 0.5$mm

	Armadillo			Bunny			Dragon			Average
	Gold	Stone	Wood	Gold	Stone	Wood	Gold	Stone	Wood	
CasMVSNet	0.8767	0.9674	**0.9685**	0.7973	0.9457	**0.9494**	0.7870	0.9001	**0.9033**	0.8995
VisMVSNet	0.6316	0.9144	0.8938	0.6709	0.9328	0.9149	0.6165	0.8791	0.8641	0.8131
COLMAP	0.7766	0.9744	0.9672	0.6303	0.9538	0.9470	0.6638	0.9085	0.8938	0.8573
DFD (SSD)	0.8393	0.4979	0.3944	0.7029	0.4583	0.3950	0.6580	0.3897	0.3012	0.5152
DFD (CNN)	**0.9219**	0.7945	0.6749	**0.8512**	0.7717	0.6776	0.8008	0.7038	0.5705	0.7519
Stereo (SSD)	0.7814	0.9747	0.9613	0.5491	0.9575	0.9140	0.6310	0.9101	0.8960	0.8417
Stereo (CNN)	0.8710	0.9748	0.9673	0.7099	**0.9581**	0.9459	0.7374	**0.9167**	0.8983	0.8866
Proposed (SSD)	0.8582	0.9748	0.9583	0.6640	0.9514	0.9093	0.7270	0.8897	0.8803	0.8681
Proposed (CNN)	0.9198	**0.9784**	0.9673	0.8324	0.9563	0.9463	**0.8259**	0.9164	0.8911	**0.9149**

Bold values indicate the top performing approach for each object and material
Results generated from data with 1% noise

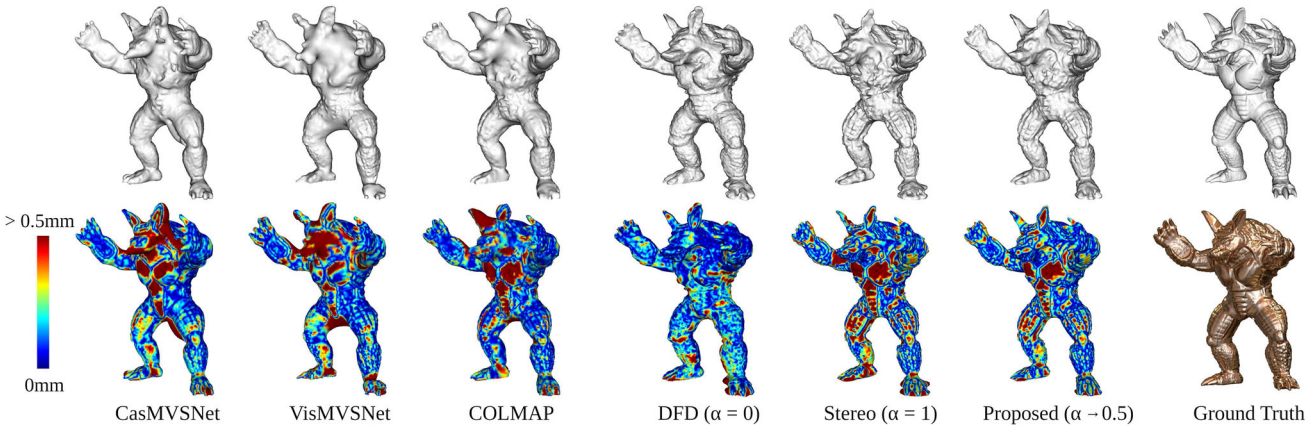

Fig. 12 Mesh reconstructions (top row) and error maps (bottom row) on the gold Armadillo dataset with 1% noise. On this dataset, defocus appears to perform best out of the comparisons shown, which we believe is due to the particularly high frequency appearance of the scratched gold material

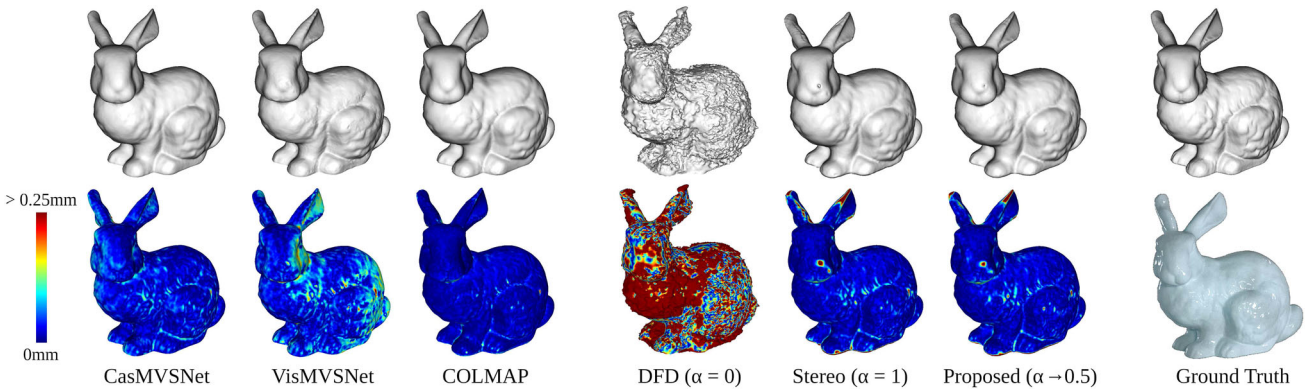

Fig. 13 Mesh reconstructions (top row) and error maps (bottom row) on the stone Bunny dataset with 1% noise. Defocus appears to fail in improving performance over our stereo term alone

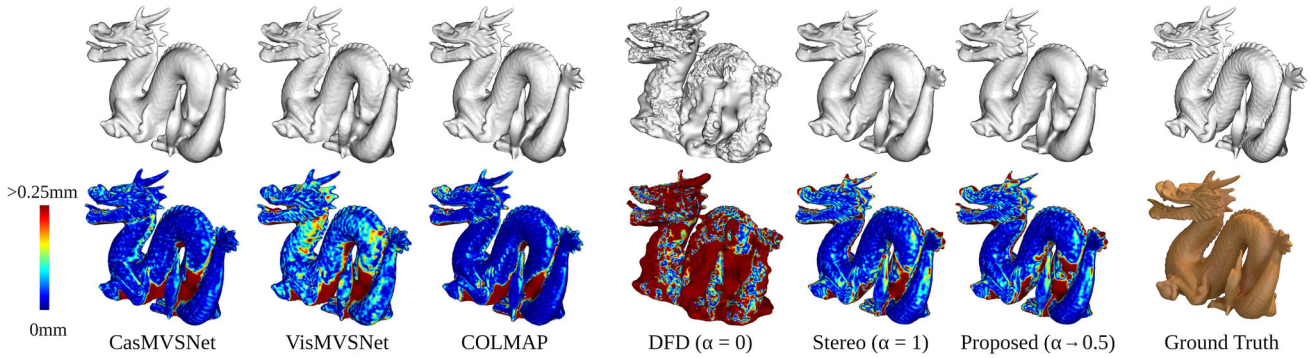

Fig. 14 Mesh reconstructions (top row) and error maps (bottom row) on the wooden Dragon dataset with 1% noise. The proposed compares well with the MVS methods

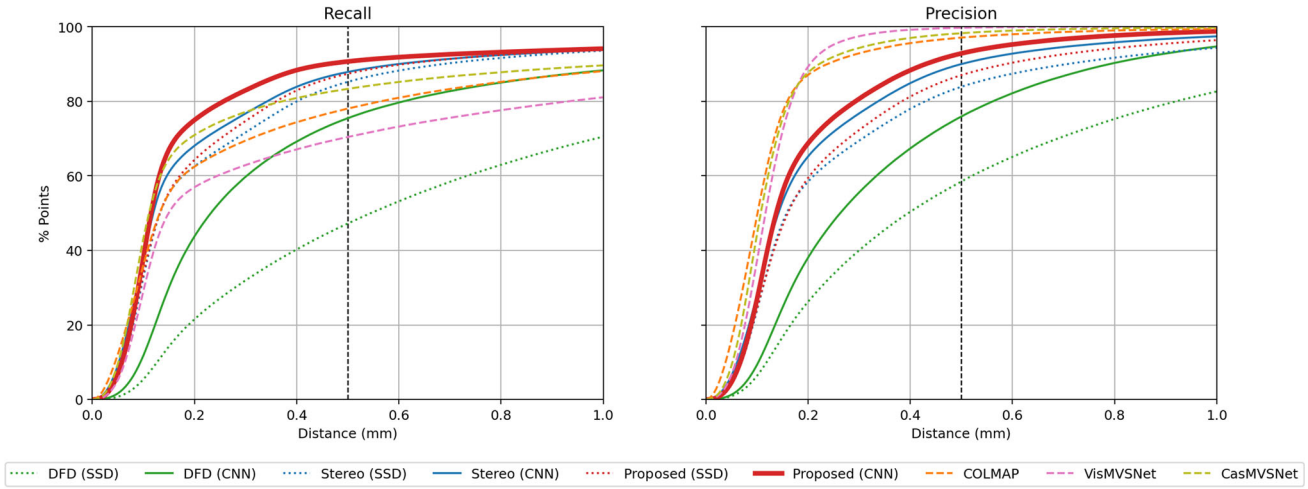

Fig. 15 Average recall (left) and precision (right) of point clouds across all synthetic experiments with 1% noise. Plots show the average percentage of points with respect to their distance from the ground truth mesh. Vertical line at 0.5mm represents the value of τ used when calculating the F-scores presented in Table 8. The proposed method with the CNN-based cost (shown in bold red) outperforms all methods by a clear margin when comparing against the recall, but trails the state-of-the-art MVS methods by around 5% in terms of precision at τ (Color figure online)

6.3.2 Real

Figures 17, 18 and 19 show a comparison of the point cloud and triangular mesh reconstructions of our real-world datasets. The Owl object is the most diffused, and so performs the most consistently across all methods. In contrast, the MVS methods struggle to achieve a complete reconstruction of the Bauble, with the meshing algorithm smoothing over holes in the point cloud. The proposed approach performs much better; achieving a complete and detailed surface.

Finally, the MVS methods fail almost completely on the Temple object due to its highly specular and reflective appearance. Our ablative study illustrates the contribution of each cue very well on this object. Defocus alone achieves a complete cloud, yet lacks finer details such as the roof ornament. Stereo alone produces a noisy cloud with many holes due to a lack of robust matches, leading to a deformed mesh. The proposed recovers a complete point cloud suitable for recovering a stable mesh complete with many details. Figure 20 illustrates this point further.

6.4 Focal Stack Ablation

Finally, we present a set of experiments that explores how the number of images in the focal stack effects the performance of the approach. For this, each variant of the method was tested with 2, 3 and 5 images of the wood Bunny dataset. When using 2 images, only the nearest and furtherest images in the focal stack are seen by the method. Note the results with 5 images are the same as those seen previously - they are presented here again for ease of comparison. As with the other synthetic experiments, 24 viewpoints are made available to the method - only the number of images per view are modified. Since DFD has been shown to tolerate noise

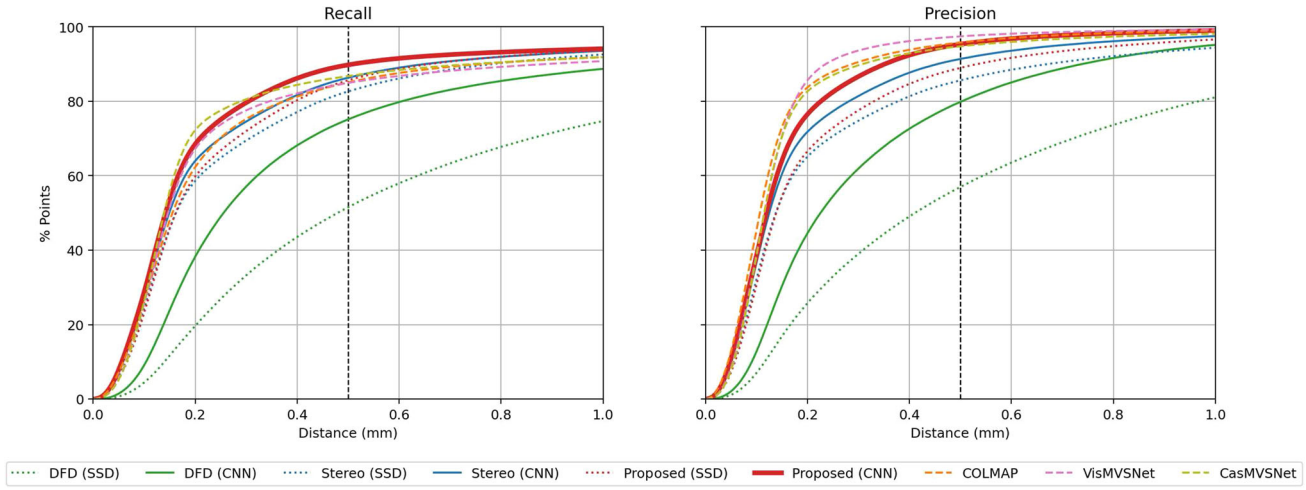

Fig. 16 Average recall (left) and precision (right) of reconstructed meshes across all synthetic experiments with 1% noise. Plots show the average percentage of vertices with respect to their distance from the ground truth mesh. For reference, the vertical line at 0.5mm represents the value of τ used when calculating the F-scores on the point clouds. The proposed method with the CNN-based cost (shown in bold red) achieves excellent recall and performs very competitively in terms of precision (Color figure online)

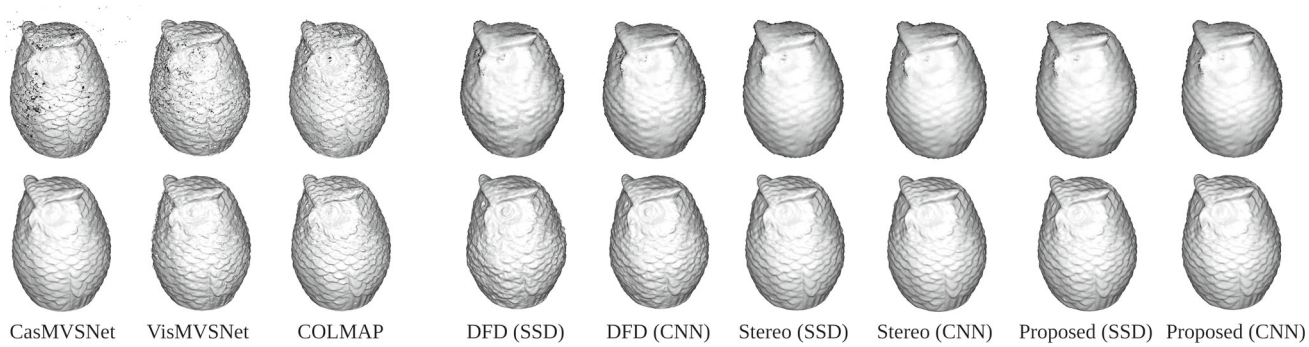

Fig. 17 3D reconstructions of the real-world Owl dataset, and a comparison of several MVS methods (left) to an ablation of the proposed method (right). Top row: filtered point clouds produced by each method. Bottom row: triangular meshes generated from Poisson surface reconstruction

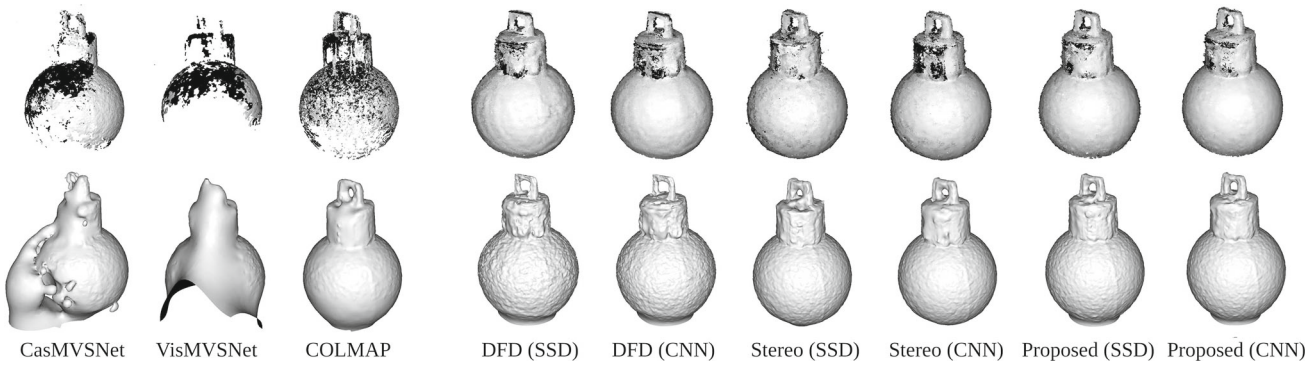

Fig. 18 3D reconstructions of the real-world Bauble dataset, and a comparison of several MVS methods (left) to an ablation of the proposed method (right). Top row: filtered point clouds produced by each method. Bottom row: triangular meshes generated from Poisson surface reconstruction

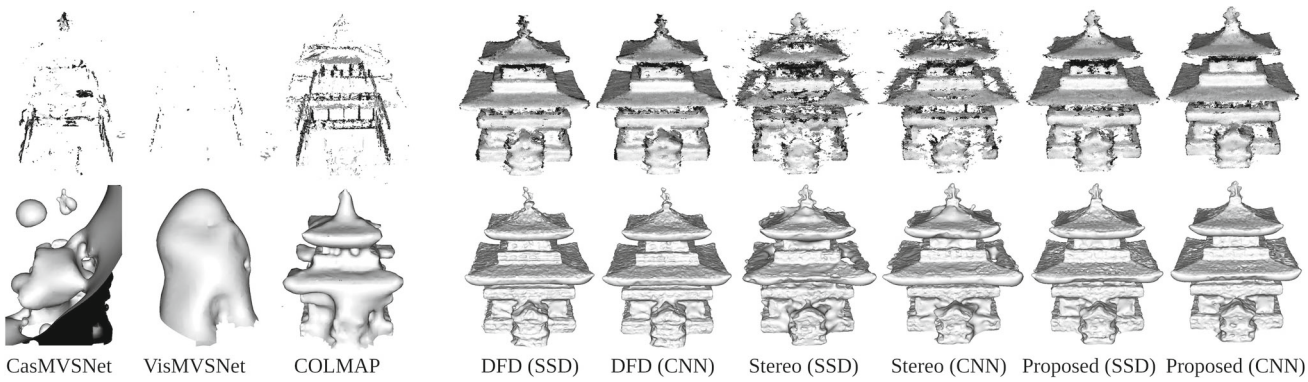

Fig. 19 3D reconstructions of the real-world Temple dataset, and a comparison of several MVS methods (left) to an ablation of the proposed method (right). Top row: filtered point clouds produced by each method. Bottom row: triangular meshes generated from Poisson surface reconstruction

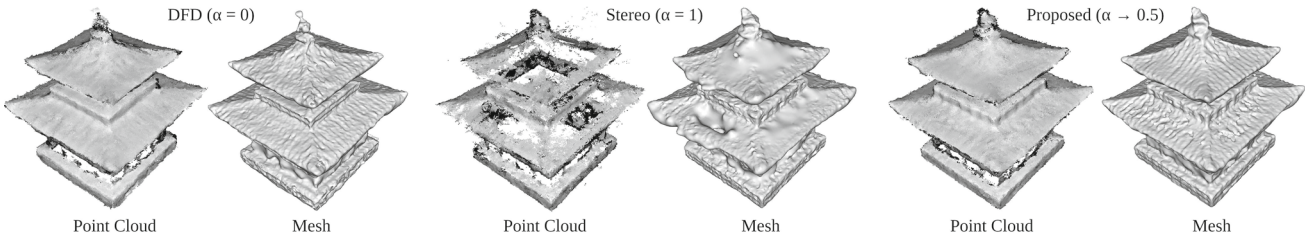

Fig. 20 Alternative view of the Temple point clouds and mesh reconstructions using the CNN cost function. Defocus alone recovers a complete cloud but lacks surface details. Stereo alone recovers a point cloud with many holes present, making it unsuitable for recovering a stable mesh. The proposed achieves the best result; recovering a complete cloud and achieving a mesh reconstruction with many surface details

Table 9 Results from ablation study where the number of images in the focal stacks are varied from 2 to 5

	Images	MSE (mm)	MAE (mm)	% Bad Pixels	Precision	Recall	F-Score
DFD (SSD)	2	1.2317	0.6655	71.33	0.7629	**0.8964**	0.8243
	3	1.2519	0.7335	71.61	0.6974	0.8650	0.7722
	5	**0.8507**	**0.5596**	**69.49**	**0.8066**	0.8955	**0.8487**
DFD (CNN)	2	0.8491	0.4603	**55.89**	0.8669	**0.9087**	0.8874
	3	0.6213	0.4718	65.71	0.8261	0.9054	0.8639
	5	**0.4898**	**0.3862**	59.14	**0.9304**	0.9056	**0.9178**
Stereo (SSD)	2	2.5127	0.7203	49.76	0.8582	0.9257	0.8906
	3	2.2331	0.6195	37.57	0.8887	0.9290	0.9084
	5	**2.0851**	**0.5751**	**33.27**	**0.8989**	**0.9310**	**0.9147**
Stereo (CNN)	2	1.7173	0.4567	30.02	0.9477	**0.9346**	0.9411
	3	1.3707	0.3627	20.55	0.9716	0.9343	0.9526
	5	**1.2247**	**0.3477**	**19.46**	**0.9734**	0.9335	**0.9530**
Both (SSD)	2	0.9976	0.4862	46.38	0.8936	0.9300	0.9114
	3	0.6833	0.3350	**29.64**	**0.9447**	**0.9316**	**0.9381**
	5	**0.6332**	**0.3256**	30.39	0.9417	0.9314	0.9365
Both (CNN)	2	0.6397	0.2777	24.80	0.9657	0.9285	0.9467
	3	0.4005	0.1932	14.92	**0.9893**	**0.9299**	**0.9587**
	5	**0.3604**	**0.1823**	**14.75**	0.9887	0.9281	0.9575

The Wood Bunny dataset with 0% noise was used to perform this study. As in the previous sections, MSE, MAE and % Bad Pixels (> 0.25mm error) are calculated from the per-viewpoint depth maps; with the value shown representing the average across all views. The F-Score metrics (Precision and Recall) are calculated from the combined point clouds, and use the ground truth mesh as reference with $\tau = 0.5$mm. Bold indicates top performer for each variant of the approach

poorly, ideal 0% noise conditions were chosen for this test.

Table 9 combines the quantitive results from the 2D depth maps (MSE, MAE and % Bad Pixels) as well as evaluation on the fused point clouds (Recall, Precision and F-score). For the 2D results, there is a clear overall improvement in MSE and MAE when more images are used. Interestingly, DFD with the CNN cost function has fewer bad pixels when using only 2 images. The reasons for this are not immediately clear, but could be related to reduced ambiguity in the DFD cost when fewer images are used.

At first glance, the results from the fused point cloud analysis appear less clear. Although half of the results indicate better performance with 5 images, the rest appear to show the opposite. On closer inspection, the majority of these conflicting values are within one thousandth of the second best performer. This indicates that the influence of the focal stack size is either marginal or generally positive depending on the metric - at least with this dataset. It is also worth noting the proposed method almost always outperforms the individual stereo and defocus terms, even with less input data. However, these results only tell some of the story, as they do not consider the influence other parameters have on reconstruction such as the aperture diameter. Nevertheless, these results verify the proposed method continues to operate coherently with smaller focal stacks.

7 Conclusion

In this paper, we have presented a complete pipeline for reconstructing scenes from multi-view finite aperture images. We began by generalising the image formation process, and introduce a novel camera calibration procedure that characterises the unavoidable formation of defocus according to thick-lens principles. Next, an MRF-based reconstruction framework was proposed that unifies defocus and stereo cues and exploits the benefits of each; achieving performance greater than the sum of its parts. In our evaluation, we demonstrate how each cue contributes to the reconstruction with an ablation study; with the proposed method exhibiting robust and consistent performance across a range of complex materials. We also explored how a feature-based cost function could benefit our reconstruction. This became even more apparent in our comparison to several MVS methods, where in most cases we achieve similar or better performance.

There are several limitations with our current approach. While our stereo term is reasonably robust and achieves performance comparable to the other MVS methods tested, the defocus term can fail under the influence of noise. Though noise is less of a concern in macro photography where large apertures and exposure times are used, it remains an unavoidable feature of the image much like defocus itself. Though the proposed method continues to work well in most cases, under adverse conditions the defocus term does appear to contribute less useful information. In future work, the modelling of noise could be introduced to improve the robustness of the defocus term to noise.

Challenges also remain concerning how best to weight the contribution of each cue. Here, a scalar weighting was used that combined the stereo and defocus cues independent to the image context; leading to residual errors where the influence of an erroneous term is particularly strong. In our experiments, this usually originated from the stereo term in low noise data, and the defocus term in high noise data. In future work, it would be interesting to introduce a contextually aware weighting, where the contribution of a cue is conditional based on the appearance of the scene. Perhaps a classifier could be implemented that can perceive in a broad sense the reflectance function of the surface, and output a weighting of cues that extracts the most performance from our framework. Finally, there are additional variables that could be explored further relating to the focal stack, such as the aperture size and number of images.

Funding This research was supported by the EPSRC (Grants EP/N 509772/1, EP/P022529/1)

Availability of data and materials The authors confirm that the datasets generated as part of this research are freely available under the terms and conditions detailed in the license agreement enclosed in the data repository. Details of the data and how to obtain access are available from the University of Surrey: https://doi.org/10.15126/surreydata.900426.

A Mutually Exclusive Solution for Pupil Ratio

Recall that our calibration defines the pupil ratio p_i as the piecewise function,

$$p_i = \begin{cases} p_1, & \text{if } p_1 < 1 \\ p_2, & \text{if } p_2 \geq 1 \end{cases}$$

where,

$$p_1 = \frac{1}{2A}\left(1 + \sqrt{1 + 4m_i A}\right), \tag{44}$$

$$p_2 = \frac{m_i}{A - 1}, \tag{45}$$

$$A = \frac{F_i}{f_\infty}\sqrt{\frac{b_\infty}{b_i}}. \tag{46}$$

Let us show that one and only one solution of p_i is valid. From Eq. 45,

$$A = \frac{m_i}{p_2} + 1. \tag{47}$$

Substituting Eq. 47 in Eq. 44,

$$p_1 = \frac{1}{2\left(\frac{m_i}{p_2} + 1\right)}\left(1 + \sqrt{1 + 4m_i\left(\frac{m_i}{p_2} + 1\right)}\right), \tag{48}$$

We need to show that p_1 is valid when p_2 is invalid. Since p_1 is valid when less than 1, we can show using Eq. 48,

$$\sqrt{1 + 4m_i\left(\frac{m_i}{p_2} + 1\right)} < \frac{2m_i}{p_2} \tag{49}$$

which simplifies to $p_2 < 1$. Similarly, when $p_1 \geq 1$ this implies $p_2 \geq 1$. Thus, only one function defining p_i gives a valid solution for all cases. This completes the proof.

References

Acharyya, A., Hudson, D., Chen, K.W., Feng, T., Kan, C.-Y., & Nguyen, T. (2016). Depth estimation from focus and disparity. In *IEEE international conference on image processing (ICIP)* (pp. 3444–3448).

Anwar, S., Hayder, Z., & Porikli, F. (2021). Deblur and deep depth from single defocus image. *Machine Vision and Applications***32**(1).

Bailey, M., & Guillemaut, J.-Y. (2020). A novel depth from defocus framework based on a thick lens camera model. In *2020 international conference on 3D vision (3DV)* (pp. 1206–1215). https://doi.org/10.1109/3DV50981.2020.00131

Bailey, M., Hilton, A., & Guillemaut, J.-Y. (2021). Finite aperture stereo: 3D reconstruction of macro-scale scenes. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2474–2484). https://doi.org/10.1109/ICCVW54120.2021.00280

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

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

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

Bradley, D., Boubekeur, T., Heidrich, W. (2008) Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *IEEE conference on computer vision and pattern recognition*, pp. 1–8.

Carvalho, M., Le Saux, B., Trouvé-Peloux, P., Almansa, A., Champagnat, F. (2019) Deep depth from defocus: How can defocus blur improve 3D estimation using dense neural networks? In *Computer Vision – ECCV 2018 Workshops* (pp. 307–323). Springer, Cham.

Chakrabarti, A., Zickler, T. (2012). Depth and deblurring from a spectrally-varying depth-of-field. In *Computer vision - ECCV. Lecture Notes in Computer Science* (pp. 648–661). Springer, Berlin, Heidelberg.

Chen, Z., Guo, X., Li, S., Cao, X., & Yu, J. (2017). A Learning-based framework for hybrid depth-from-defocus and stereo matching. arXiv e-prints arXiv:1708.00583 [cs.CV]

Chen, R., Han, S., Xu, J., Su, H. (2019) Point-based multi-view stereo network. CoRR abs/1908.04422. arXiv:1908.04422.

Chen, L. Y., Shuochen, S., & S., Matsushita, S., KunZhou, S.,& Lin, S. (2016). Bayesian depth-from-defocus with shading constraints. *IEEE Transactions on Image Processing,25*(2), 589–600.

Chen, C.-H., Zhou, H., & Ahonen, T. (2015). Blur-Aware Disparity Estimation from Defocus stereo images. In *2015 IEEE international conference on computer vision (ICCV)* (Vol. 2015, pp. 855–863).

Choy, C.B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016). 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*.

Delaunoy, A., Pollefeys, M. (2014). Photometric bundle adjustment for dense multi-view 3d modeling. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 1486–1493). https://doi.org/10.1109/CVPR.2014.193.

Emerson, D.R., & Christopher, L.A. (2019). 3-D scene reconstruction using depth from defocus and deep learning. In *2019 IEEE applied imagery pattern recognition workshop (AIPR)* (pp. 1–8). https://doi.org/10.1109/AIPR47015.2019.9174568

Favaro, P. (2007). Shape from focus and defocus: Convexity, quasiconvexity and defocus-invariant textures. In *IEEE 11th international conference on computer vision (ICCV)* (pp. 1–7).

Favaro, P. (2010). Recovering thin structures via nonlocal-means regularization with application to depth from defocus. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1133–1140).

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

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

Furukawa, Y., & Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *IEEE Transactions On Pattern Analysis And Machine Intelligence, 32*(8), 1362–1376.

Gheţa, I., Frese, C., Heizmann, M., & Beyerer, J. (2007). A new approach for estimating depth by fusing stereo and defocus information. In R. Koschke, O. Herzog, K.-H. Rödiger, & M. Ronthaler (Eds.), *Informatik 2007 - Informatik Trifft Logistik* - (Vol. 1, pp. 26–31). Bonn: Gesellschaft für Informatik e. V.

Gu, X., Fan, Z., Dai, Z., Zhu, S., Tan, F., & Tan, P. (2019). Cascade cost volume for high-resolution multi-view stereo and stereo matching (pp. 1912–06378).

Hartley, R. (2000). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hasinoff, S., & Kutulakos, K. (2009). Confocal stereo. *International Journal of Computer Vision, 81*(1), 82–104.

Hornung, A., Kobbelt, L. (2006). Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *IEEE computer society conference on computer*

vision and pattern recognition (CVPR) (Vol. 1, pp. 503–510). https://doi.org/10.1109/CVPR.2006.135

Huang, P., Matzen, K., Kopf, J., Ahuja, N., & Huang, J. (2018). DeepMVS: Learning multi-view stereopsis. CoRR arXiv:1804.00650.

Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L. (2017). Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 2307–2315).

Kar, A., Häne, C., Malik, J. (2017) Learning a multi-view stereo machine.

Kashiwagi, M., Mishima, N., Kozakaya, T., & Hiura, S. (2019). Deep depth from aberration map. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 4069–4078). https://doi.org/10.1109/ICCV.2019.00417

Kazhdan, M., & Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Transactions on Graphics, 32*(3), 1–13.

Kingslake, R. (1992). Optics in photography. SPIE Press monograph; PM06. SPIE, Bellingham, Wash. (1000 20th St. Bellingham WA 98225-6705 USA)

Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics***36**(4).

Kuhn, A., Sormann, C., Rossi, M., Erdler, O., & Fraundorfer, F. (2020). DeepC-MVS: deep confidence prediction for multi-view stereo reconstruction. https://doi.org/10.1109/3DV50981.2020.00050

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

Li, Z., Wang, K., Zuo, W., Meng, D., & Zhang, L. (2016). Detail-preserving and content-aware variational multi-view stereo reconstruction. *IEEE Transactions on Image Processing***25**(2).

Li, G., & Zucker, S. W. (2010). Differential geometric inference in surface stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(1), 72–86.

Lin, H., Chen, C., Kang, S.B., & Yu, J. (2015). Depth recovery from light field using focal stack symmetry. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 3451–3459). https://doi.org/10.1109/ICCV.2015.394

Lin, X., Suo, J., Cao, X., & Dai, Q. (2013). Iterative feedback estimation of depth and radiance from defocused images. In *Lecture Notes in Computer Science 11th Asian conference on computer vision (ACCV)* (Vol. 7727, pp. 95–109). Berlin, Heidelberg: Springer.

Liu, Y., Dai, Q., & Xu, W. (2010). A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Transactions on Visualization and Computer Graphics, 16*(3), 407–418.

Logothetis, F., Mecca, R., & Cipolla, R. (2019). A differential volumetric approach to multi-view photometric stereo. In *IEEE/CVF international conference on computer vision (ICCV)* (pp. 1052–1061). https://doi.org/10.1109/ICCV.2019.00114

Luo, K., Guan, T., Ju, L., Huang, H., Luo, Y. (2019). P-MVSNet: Learning Patch-Wise Matching Confidence Aggregation for Multi-View Stereo. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 10451–10460). https://doi.org/10.1109/ICCV.2019.01055

Mannan, F., Langer, M.S. (2015). Optimal camera parameters for depth from defocus. In *International conference on 3D vision* (pp. 326–334).

Martinello, M., Wajs, A., Quan, S., Lee, H., Lim, C., Woo, T., Lee, W., Kim, S.-S., & Lee, D. (2015) Dual aperture photography: Image and depth from a mobile camera. In *IEEE international conference on computational photography (ICCP)* (pp. 1–10).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision - ECCV 2020* (pp. 405–421). Cham: Springer.

Moeller, M., Benning, M., Schonlieb, C., & Cremers, D. (2015). Variational depth from focus reconstruction. *IEEE Transactions on Image Processing, 24*(12), 5369–5378.

Namboodiri, V.P., Chaudhuri, S., & Hadap, S. (2008). Regularized depth from defocus. In *15th IEEE international conference on image processing (ICIP)* (pp. 1520–1523).

Nayar, S. K., & Nakagawa, Y. (1994). Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(8), 824–831. https://doi.org/10.1109/34.308479.

Olsson, C., Ulén, J., Boykov, Y. (2013). In defense of 3D-label stereo. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1730–1737).

Paramonov, V., Panchenko, I., Bucha, V., Drogolyub, A., & Zagoruyko, S. (2016). Depth camera based on color-coded aperture. In *2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 910–918).https://doi.org/10.1109/CVPRW.2016.118

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

Persch, N., Schroers, C., Setzer, S., & Weickert, J. (2017). Physically inspired depth-from-defocus. *Image and Vision Computing, 57,* 114–129.

Rajagopalan, A.N., Chaudhuri, S., & Mudenagudi, U. (2004). Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence***26**(11).

Rother, C., Kolmogorov, V., & Blake, A. (2004). *GrabCut: Interactive foreground extraction using iterated graph cuts*. New York, NY, USA: Association for Computing Machinery.

Rowlands, A. (2017). Fundamental optical formulae. *Physics of Digital Photography*. https://doi.org/10.1088/978-0-7503-1242-4ch1.

Schönberger, J. L., Zheng, E., Frahm, J.-M., & Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision - ECCV 2016* (pp. 501–518). Cham: Springer.

Song, G., & Lee, K.M. (2018) Depth estimation network for dual defocused images with different depth-of-field. In *25th IEEE international conference on image processing (ICIP)* (pp. 1563–1567).

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

Takeda, Y., Hiura, S., & Sato, K. (2013). Fusing depth from defocus and stereo with coded apertures. In *2013 IEEE conference on computer vision and pattern recognition* (pp. 209–216). https://doi.org/10.1109/CVPR.2013.34

Tang, H., Cohen, S., Price, B., Schiller, S., & Kutulakos, K. N. (2017). Depth from defocus in the wild. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4773–4781.

Tao, M. W., Hadap, S., Malik, J., & Ramamoorthi, R. (2013). Depth from combining defocus and correspondence using light-field cameras. In *IEEE international conference on computer vision* (pp. 673–680).

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

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

Tola, E., Strecha, C., & Fua, P. (2012). Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications, 23*(5), 903–920.

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

Wang, T.-C., Srikanth, M., & Ramamoorthi, R. (2016). Depth from semi-calibrated stereo and defocus. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3717–3726).

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

Wu, C., Wilburn, B., Matsushita, Y., & Theobalt, C. (2011). High-quality shape from multi-view stereo and shading under general illumination. In *CVPR 2011* (pp. 969–976).

Yao, Y., Luo, Z., Li, S., Fang, T.,& Quan, L. (2018). MVSNet: Depth inference for unstructured multi-view stereo. In *Computer Vision–ECCV 2018* (pp. 785–801). Springer, Cham.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5520–5529). https://doi.org/10.1109/CVPR.2019.00567.

Zagoruyko, S., Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *IEEE conference on computer vision and pattern recognition. Proceedings* (Vol. 07-12, pp. 43534361– (2015-06-01)). http://search.proquest.com/docview/1770338817/.

Zhang, J., Yao, Y., Li, S., Luo, Z., & Fang, T. (2020). Visibility-aware multi-view stereo network. arXiv e-prints arXiv:2008.07928 [cs.CV].

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

Zhu, Z., Stamatopoulos, C., & Fraser, C. S. (2015). Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing, 109*(C), 47–61.