International Journal of Computer Vision (2023) 131:2243–2266
https://doi.org/10.1007/s11263-023-01795-w

A Family of Approaches for Full 3D Reconstruction of Objects with Complex Surface Reflectance

Gianmarco Addari[1] · **Jean-Yves Guillemaut**[1]

Received: 2 March 2022 / Accepted: 7 April 2023 / Published online: 1 June 2023

Abstract
3D reconstruction of general scenes remains an open challenge with current techniques often reliant on assumptions on the scene's surface reflectance, which restrict the range of objects that can be modelled. Helmholtz Stereopsis offers an appealing framework to make the modelling process agnostic to surface reflectance. However, previous formulations have been almost exclusively limited to 2.5D modelling. To address this gap, this paper introduces a family of reconstruction approaches that exploit Helmholtz reciprocity to produce complete 3D models of objects with arbitrary unknown reflectance. This includes an approach based on the fusion of (orthographic or perspective) view-dependent reconstructions, a volumetric approach optimising surface location within a voxel grid, and a mesh-based formulation optimising vertices positions of a given mesh topology. The contributed approaches are evaluated on synthetic and real datasets, including novel full 3D datasets publicly released with this paper, with experimental comparison against a wide range of competing methods. Results demonstrate the benefits of the different approaches and their abilities to achieve high quality full 3D reconstructions of complex objects.

Keywords Helmholtz stereopsis · 3D reconstruction · Complex surface reflectance · Markov Random Fields

1 Introduction

Major advances have been made in scene modelling over the past decades through the development of both classical and more recently deep learning approaches. These leverage different cues and features, whether handcrafted or learnt, to infer scene geometry, often with an impressive degree of fidelity. However, existing techniques usually remain reliant on assumptions on the scene properties (surface reflectance, texture, geometry) or capture conditions (illumination, camera placement) to produce reliable reconstructions. Surface reflectance in particular is one of the main factors that currently prevents generalisation of modelling techniques to arbitrary scenes as it is often assumed to either follow a particular model (e.g. Lambertian, or specific parametric model) or be known a priori. This restricts the applicability of existing techniques to specific classes of objects which meet those requirements.

Recent developments in neural radiance fields (Mildenhall et al., 2021; Wang et al., 2021; Yariv et al., 2021; Oechsle et al., 2021) have relaxed the requirement to have a model of the Bidirectional Reflectance Distribution Function (BRDF). Another promising approach that makes the reconstruction process agnostic to the scene's reflectance is Helmholtz Stereopsis (HS). The approach exploits Helmholtz reciprocity (von Helmholtz, 1924) to derive a constraint that is independent of the BRDF and can be used to retrieve both scene depth and normal information. Whilst promising, its formulations to date have been almost exclusively limited to 2.5D reconstruction, framing the problem in terms of estimating the depth and normals at each pixel from a given viewpoint (usually a virtual view) and thereby allowing only a partial reconstruction. These approaches also ignore visibility since all the cameras are confined to one side of the object. Aside from our recent work which we extend here, the few HS approaches that considered full 3D reconstruction were either based on local optimisation or refinement in conjunction with another modality (structured light).

This paper aims to fill the gap in this area by proposing a family of methods for full 3D modelling of scenes

Communicated by Andrea Fusiello.

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

Gianmarco Addari
g.addari@surrey.ac.uk

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

with arbitrary unknown reflectance based on *Helmholtz reciprocity*. It explores different paradigms to extract a full 3D model from a set of reciprocal image pairs each acquired by swapping the positions of the camera and light source. The first approach is based on fusing view-dependent reconstructions, leveraging confidence metrics to optimise the full 3D surface recovery. Different formulations are investigated depending on the geometry of the grid used to recover the intermediate view-dependent representations (orthographic or perspective). The second approach is based on a volumetric optimisation process which overcomes the need to compute intermediate representations. The last approach uses a mesh-based formulation to also allow direct optimisation of the full 3D surface while at the same time reducing the computational footprint compared to the volumetric formulation.

The paper makes the following contributions. Firstly, it introduces three novel approaches to perform full 3D modelling of scenes with complex reflectance. The formulations share the use of Markov Random Fields (MRFs) to provide a principled optimisation framework. A tailored visibility handling approach is also introduced to overcome the shortcomings of previous 2.5D formulations. Secondly, the paper contributes novel datasets for full 3D reconstruction using HS. These include both synthetic scenes generated using POV-Ray and real scenes acquired using a versatile capture setup built from consumer hardware (a pair of DSLR cameras equipped with lens-mounted flashes). The datasets are publicly released with this paper (see data access statement at the end of the paper). These datasets alongside other publicly available datasets are used to conduct an extensive experimental validation of the proposed approaches including a comparison against recent Multi-View Stereo (MVS), MultiView Photometric Stereo (MVPS) and neural radiance field-based methods.

The paper extends our previous work presented in Addari and Guillemaut (2019a, b, 2020) in several ways. First, all three paradigms are brought together into a common article with full detail provided and a common evaluation protocol to facilitate their analysis and comparison. Second, the fused view-dependent formulation is generalised to perspective cameras and confidence measures are introduced. The former allows segmentation information to be leveraged during the modelling process while the latter improves the fusion process. Third, a substantially expanded experimental validation is conducted. This incorporates evaluation on a large number of scenes as well as a thorough analysis using additional performance metrics, new results and for the first time a comparison against the state of the art in MVS, MVPS and neural radiance field-based methods to validate experimentally the benefits of HS for reconstruction of surfaces with complex reflectance.

The paper is structured as follows. Section 2 reviews the main 3D reconstruction approaches, highlighting their dependency on the scene's reflectance properties as well as the state of the art in HS and its underpinning principles. Section 3 describes the three proposed approaches to full 3D modelling of complex scenes. Section 4 conducts an experimental evaluation of the different approaches on both synthetic and real data. Finally, Sect. 5 concludes the paper and discusses avenues for future work.

2 Related Work

2.1 3D Reconstruction Overview

This section provides an overview of the main categories of 3D reconstruction techniques, with an emphasis on their abilities to handle different types of surface reflectance properties.

Shape from Silhouettes (SfS) approaches are based on intersecting the set of visual cones defined by backprojecting the object's silhouette in each image (Baumgart, 1974) to obtain its Visual Hull (VH) (Laurentini, 1994). Independence from the surface properties is the main advantage offered by this class of methods as long as good background segmentation can be obtained. SfS methods can be further divided into image-based (Matusik et al., 2000), volumetric (Szeliski, 1993; Tarini et al., 2002; Liu et al., 2006), surface-based (Cipolla & Blake, 1992; Forbes et al., 2004; Liang & Wong, 2010) and hybrid approaches (Boyer & Franco, 2003). Further improvements have been introduced by using Convolutional Neural Networks (CNNs) to obtain a probabilistic VH (Gilbert et al., 2018). However, this class of techniques remains inherently limited to the reconstruction of convex objects due to the impossibility of visualising concavities in silhouettes.

Binocular Stereo and Multi-View Stereo (MVS) methods (Szeliski et al., 2008; Seitz et al., 2006) use point correspondences across images to infer surface depth. Contrary to SfS they are not limited to the reconstruction of convex objects. However, the reconstructed surfaces are often assumed to be Lambertian or sufficiently textured to perform point matching across views. This assumption is the main drawback as it is often violated in practice. Certain approaches attempt to jointly estimate surface reflectance and scene geometry. In Oxholm and Nishino (2014) and Lombardi and Nishino (2016), the geometry is obtained iteratively form a previous estimate of the reflectance and vice versa. This hinders performance, as each solution will only be as good as the previously estimated other term. In Holroyd et al. (2010), instead, light descattering is used in conjunction with an active MVS technique to obtain both geometry and BRDF. The results are highly precise, however the method requires a complex setup comprising two coaxial camera/light source assemblies to capture stacks of coaxial images and reciprocal

images of the object. More recently, several MVS formulations have explored the use of CNNs. Examples include (Kar et al., 2015), in which reconstruction is limited to a specific set of categories, and Choy et al. (2016), where volumetric rendition of the analysed scenes is class-independent. Despite the potential of CNN-based methods, they have yet to reach the level of accuracy of other types of approaches when a larger number of images are used. A further hindrance that is still holding back machine learning approaches in this context is the availability of training data. Recently unsupervised solutions have been proposed such as in Dai et al. (2019), however the quality of results is still far from that of traditional techniques.

Photometric Stereo (PS) (Woodham, 1980) allows reconstruction of non-Lambertian surfaces, however the surface reflectance needs to be known a priori. An extensive survey of PS techniques was published in Ackermann and Goesele (2015). Han and Shen (2015) use a complex setup to obtain a very densely sampled set of lights and viewing directions. This allows to exploit specularities and shadows to obtain the reconstruction of both isotropic and anisotropic objects. The BRDF is modelled by dividing it into its components: diffuse, specular and shadows. Capturing many images under varying lighting allows to perform BRDF modelling and PS reconstruction. The main drawbacks of this approach are its computational complexity and the unsatisfactory accuracy obtained when few lighting directions are used. In Ikehata (2018), a CNN-based approach is proposed to reconstruct non-convex objects. The approach learns relations between the reflections in the input images and surface normal orientation, using synthetic datasets for training. A similar work is applied to near-field PS in Logothetis et al. (2020). Despite showing promising results, the surface estimation remains tied to the per point network predictions and no explicit surface optimisation strategy is implemented. PS removes the need to establish correspondences across views required in stereo approaches, allowing for scenes with more complex reflectance to be reconstructed. Further, it achieves high quality reconstruction under Lambertian conditions or when an accurate model of the scene reflectance is available. However, the Lambertian assumption is often violated and it remains challenging to obtain an accurate model of the scene reflectance (Ward, 1992; Tunwattanapong et al., 2013). Many challenges remain in achieving accurate reconstruction of non-Lambertian surfaces which present spatially varying or anisotropic BRDF and most PS approaches still rely on the assumption of isotropic reflectance to operate. Multi-View Photometric Stereo (MVPS) methods generalise PS to multiple viewpoints enabling both geometric and photometric cues to be leveraged in the same framework (Park et al., 2016; Logothetis et al., 2019; Li et al., 2020).

Another important set of techniques used to perform 3D reconstruction is based on separating specular and diffuse reflections. For example, in Mallick et al. (2005), a data-driven colour space conversion is performed on the RGB images, allowing to separate the two components. The authors present this method for surfaces that can be modelled using dichromatic reflectance, a special case of BRDF. Performing this separation allows to only consider the two-channel diffuse component, which can be approximated as Lambertian, to produce accurate reconstructions using PS. Similarly, Ma et al. (2007) propose the use of spherical gradient illumination patterns to separate the specular and diffuse components of objects made up of complex materials such as human faces. The main drawbacks of this technique are the single viewpoint and the trade-off at the edge of the objects. Conversely, in Ghosh et al. (2011), the authors use two polarised spherical gradient illumination patterns to perform the reconstruction with multiple viewpoints. The cameras are equipped with polarisers to selectively capture the correct pattern. Finally, in Fyffe et al. (2016) the authors perform colour space conversion and use the results to obtain the diffuse normals and albedo and the specular normals for each colour sub-space. They further compute a per-pixel specular exponent to refine the resulting mesh. An advantage of this method with respect to Ma et al. (2007) and Ghosh et al. (2011) is that off-the-shelf components are used and the total capture time is significantly lower. However the setup is still fairly complex and requires a large number of cameras and flashes.

Recent advances in neural scene representations and volume rendering techniques have enabled an unprecedented level of photorealism in novel view synthesis from only a sparse number of input views, pioneered by the introduction of NeRF (Mildenhall et al., 2021). The approach is based on combining a neural radiance field scene representation, which uses a neural network to model both radiance and volume density at each point in space, with a volume rendering approach which, being differentiable, is particularly well suited for learning the network's weights. Although the approach is primarily aimed at novel view synthesis, scene geometry can be retrieved from the volume density. However, the quality of the surface that can be extracted is limited since the network is tailored for novel view synthesis rather than reconstruction. Several concurrent subsequent works have extended the approach to address this limitation and enable high-quality surface modelling by incorporating an implicit surface representation into the framework (Wang et al., 2021; Yariv et al., 2021; Oechsle et al., 2021). Wang et al. (2021) and Yariv et al. (2021) achieve this by parametrising the volume density based on a signed distance function, while Oechsle et al. (2021) introduce a continuous occupancy field to represent the surface. All approaches demonstrate the ability to recover fine detail including thin structures. Interestingly, these techniques do not make any assumption about the scene reflectance. Other approaches

combine inverse rendering techniques with deep learning to recover shape alongside material properties (Bi et al., 2020; Zhang et al., 2021). Bi et al. (2020) use a volumetric representation to model opacity, surface normal and reflectance at each voxel, and they require the images to be acquired under the assumption of a collocated light source. Zhang et al. (2021) use a signed-distance function to represent shape and spherical Gaussians to approximate the light transport. Their approach is able to infer the scene properties from a set of images captured under static illumination.

The vast majority of existing methods, aside from some of the more recent learning-based approaches using neural radiance fields, fail to address surfaces presenting complex and varying reflectance as they mainly rely on simplifying assumptions to constrain the problem. An interesting alternative to these techniques that allows reconstruction of surfaces with complex BRDFs is HS.

2.2 Helmholtz Stereopsis

Helmholtz Stereopsis (HS) exploits the principle of Helmholtz reciprocity (von Helmholtz, 1924) to make the reconstruction process agnostic to the reflectance. This was first introduced in Magda et al. (2001) and subsequently developed into what is known as HS in Zickler et al. (2002). The key idea is to exploit Helmholtz reciprocity as a constraint to identify 3D points located on the surface of an object and their corresponding normals. Helmholtz reciprocity states that the measured BRDF at a surface point remains invariant when illumination and viewing directions are swapped (Nicodemus et al., 1977; Snyder, 2002). This invariance is exploited to obtain the normal of a surface point, using multiple reciprocal pairs of images from which the point is visible. In its original formulation, HS is posed as a maximum likelihood problem, with no regularisation, which can yield noisy results.

Given a point light source with strength κ located at $\boldsymbol{O}_\mathrm{r}$, the intensity of the surface point $\boldsymbol{P}$ imaged by a camera positioned at $\boldsymbol{O}_\mathrm{l}$ satisfies:

$$i_\mathrm{l} = f(\boldsymbol{v}_\mathrm{r}, \boldsymbol{v}_\mathrm{l}) \frac{\boldsymbol{n} \cdot \boldsymbol{v}_\mathrm{r}}{\|\boldsymbol{O}_\mathrm{r} - \boldsymbol{P}\|^2} \kappa, \tag{1}$$

where $f(\boldsymbol{v}_\mathrm{r}, \boldsymbol{v}_\mathrm{l})$ is the BRDF at surface point $\boldsymbol{P}$ (see Fig. 1). The unit vectors $\boldsymbol{v}_\mathrm{r}$ and $\boldsymbol{v}_\mathrm{l}$ indicate the lighting and viewing directions respectively, while $\boldsymbol{n}$ denotes the surface normal at point $\boldsymbol{P}$. When inverting camera and light positions, the same point $\boldsymbol{P}$ viewed by the camera located at $\boldsymbol{O}_\mathrm{r}$ and lit by a light source with strength κ located at $\boldsymbol{O}_\mathrm{l}$ satisfies:

$$i_\mathrm{r} = f(\boldsymbol{v}_\mathrm{l}, \boldsymbol{v}_\mathrm{r}) \frac{\boldsymbol{n} \cdot \boldsymbol{v}_\mathrm{l}}{\|\boldsymbol{O}_\mathrm{l} - \boldsymbol{P}\|^2} \kappa. \tag{2}$$

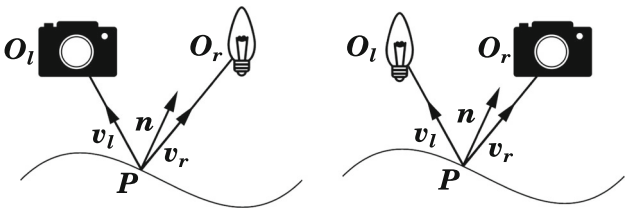

Fig. 1 Helmholtz Stereopsis (HS) acquisition principle. A reciprocal pair of images is taken by swapping camera and light positions ($\boldsymbol{O}_\mathrm{l}$ and $\boldsymbol{O}_\mathrm{r}$). The BRDF at a surface point ($\boldsymbol{P}$) is invariant in the two scenarios, allowing to obtain its normal ($\boldsymbol{n}$) when at least three reciprocal pairs are considered from different positions

Helmholtz reciprocity states that the reflectance measured at point $\boldsymbol{P}$ remains unchanged when the positions of cameras and light source are interchanged, i.e. $f(\boldsymbol{v}_\mathrm{r}, \boldsymbol{v}_\mathrm{l}) = f(\boldsymbol{v}_\mathrm{l}, \boldsymbol{v}_\mathrm{r})$, from which the following BRDF-agnostic constraint is obtained:

$$\left(i_\mathrm{l} \frac{\boldsymbol{v}_\mathrm{l}}{\|\boldsymbol{O}_\mathrm{l} - \boldsymbol{P}\|^2} - i_\mathrm{r} \frac{\boldsymbol{v}_\mathrm{r}}{\|\boldsymbol{O}_\mathrm{r} - \boldsymbol{P}\|^2} \right) \cdot \boldsymbol{n} = \boldsymbol{w} \cdot \boldsymbol{n} = 0. \tag{3}$$

With three or more camera pairs, both the surface location and its normal can be recovered. Stacking all vectors $\boldsymbol{w}$ into a matrix W and performing Singular Value Decomposition (SVD), the ratio of the second and third singular values σ_2 and σ_3 can be used to define a measure of the coplanarity of the $\boldsymbol{w}$ vectors which, in turn, provides a means to identify surface points. Once the surface location has been identified, the last singular vector of the corresponding point provides an estimate of the surface normal.

To relax the requirement of having at least three camera/light pairs, binocular formulations (Tu & Mendonca, 2003; Zickler et al., 2003) were subsequently introduced, employing a single pair of reciprocal images for reconstruction. Both works utilise a partial differential equation to compute the surface depth at each epipolar line, obtaining a set of solutions that are then disambiguated with further optimisation. In Guillemaut et al. (2004) the authors extend the classes of surface that can be reconstructed to strongly textured and rough ones, by performing HS over image patches instead of using single pixels. In Janko et al. (2004), Janko et al. propose a radiometric calibration approach to account for variations in pixel sensitivity and non-isotropic illumination. In Zickler (2006), specular highlights, which have precise correspondences across reciprocal pairs, are used to perform radiometric and geometric calibration, without the need of acquiring additional images. Finally, in Guillemaut et al. (2008) a different normal error measure is proposed in the form of the radiometric distance function.

An alternative to maximum likelihood for classic HS was proposed in Roubtsova and Guillemaut (2014a, b, 2017, 2018). They present a maximum a posteriori formulation for both classic HS (Roubtsova & Guillemaut, 2014a, 2018)

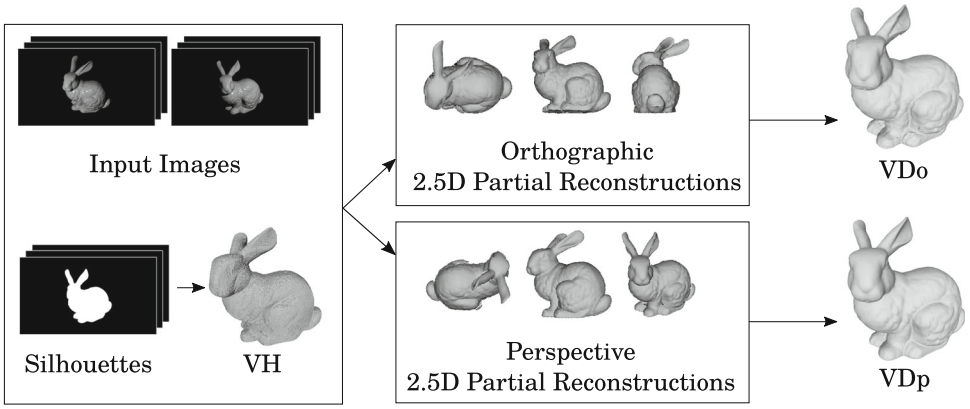

Fig. 2 Pipeline for the fused view-dependent HS approaches. The methods are initialised using a Visual Hull (VH) obtained from silhouettes of the input images. Multiple partial reconstructions are then obtained using either orthographic sampling (VDo approach) or perspective sampling (VDp approach). The partial surfaces are then integrated using confidence scores and Poisson surface reconstruction

and colour HS (Roubtsova & Guillemaut, 2014b, 2017). In the latter, wavelength multiplexing is used to extend HS to dynamic surfaces. All the aforementioned techniques are applied to 2.5D surfaces and do not handle visibility or occlusions directly. Furthermore, they all use virtual orthographic cameras to perform the reconstruction, which does not fully leverage the available segmentation information and can cause artefacts when the methods attempt to estimate depth at pixels where the object's surface may not be visible.

In Weinmann et al. (2012), HS is used in conjunction with a structured light approach as a refinement step to perform full 3D reconstruction. The technique is employed on areas of the surface where fine details are present, while the low frequency shape of the object is obtained using structured light. A complex setup consisting of a light dome is used, making this method difficult to reproduce and constrained to a limited set of scenes. The first time HS is used on its own to perform full 3D reconstruction is in Delaunoy et al. (2010), where gradient descent is proposed to perform the optimisation. The faces of the initial surface are iteratively moved towards a lower energy solution. Despite obtaining convincing results, the use of gradient descent makes the method prone to local minima and does not guarantee a globally optimal solution.

To date, most HS approaches have been limited to performing 2.5D reconstructions. In contrast, this paper introduces several methods to achieve full 3D reconstruction of scenes with arbitrary unknown reflectance, exploring different strategies to improve performance. This paper presents our previous work from Addari and Guillemaut (2019a, 2020, 2019b) in a unified light and extends it in several ways. Firstly, this advances our earlier formulation by generalising the fused view-dependent approach to full perspective enabling us to leverage segmentation information and improve modelling accuracy. This generalisation is shown to translate into a significant improvement in performance compared to our previous orthographic formulation. We also introduce several confidence metrics to improve the fusion of view-dependent reconstructions. Second, the experimental evaluation has been substantially expanded in terms of number of datasets considered, the depth of the analysis which now incorporates a quantitative evaluation for each dataset, and some additional benchmarking results on the DiLiGenT-MV dataset. The proposed approaches are evaluated against a wide range of approaches including two MVS methods, one neural radiance field-based approach and two MVPS methods. The final contribution is the release of datasets to support further research in the field and allow benchmarking of algorithms for reconstruction of scenes with complex surface reflectance.

3 Methodology

3.1 Fused View-Dependent Helmholtz Stereopsis

In this first approach, partial reconstructions are performed from multiple viewpoints around the objects which are then fused to obtain a full 3D model (see illustration of pipeline in Fig. 2). Different approaches are possible depending on the types and placement of the viewpoints used to perform the partial reconstructions. In particular, two variants are considered. The first one consists in using virtual orthographic cameras. In practice, six orthographic views are considered, coinciding with each of the three world axes, with two directions per axis. This provides an intuitive way of sampling the scene volume allowing a full coverage of the object from a minimal number of reconstructions. The second variant is based on using perspective views. In this case, each partial reconstruction can be performed directly from the viewpoint of one of the input cameras used to acquire the scene. This second approach presents the benefit of being able to leverage segmentation information during the reconstruction process by ensuring that all pixels for which a depth is estimated correspond to the object (something that is not feasible in the case of an orthographic reconstruction where no image has been acquired from the reconstruction viewpoint).

More formally, given a set of reciprocal pairs of images $\mathcal{I}_0^{\mathrm{l}}, \mathcal{I}_0^{\mathrm{r}}, \ldots, \mathcal{I}_{n-1}^{\mathrm{l}}, \mathcal{I}_{n-1}^{\mathrm{r}}$ and a camera from which reconstruc-

tion is performed, a 2-dimensional grid $\mathcal{G}$ is created over the image plane, where each node directly corresponds to a pixel in the frame of the camera used for reconstruction. Each node, or pixel, will be assigned a label indicating at what depth the surface is found when backprojecting the image point. The depth labels are denoted by: $d_0, \ldots, d_{N-1}$, where d_0 and d_{N-1} correspond to the closest and furthest points respectively.

To estimate the depth at which the surface is located at each pixel of the camera used for reconstruction, an MRF is constructed. Each node corresponds to an entry in the grid and an energy function, containing a data term to express the surface occupancy likelihood and a smoothness term to regularise the solution across all nodes, is defined. The energy function is formulated as follows:

$$E(\boldsymbol{d}) = (1-\alpha)\sum_{\boldsymbol{p}\in\mathcal{G}} D_{2\mathrm{D}}(B(\boldsymbol{p}, d_{\boldsymbol{p}})) + \alpha \sum_{(\boldsymbol{p},\boldsymbol{q})\in\mathcal{N}_{2\mathrm{D}}} S_{2\mathrm{D}}(B(\boldsymbol{p}, d_{\boldsymbol{p}}), B(\boldsymbol{q}, d_{\boldsymbol{q}})), \tag{4}$$

where α is a balancing parameter between the data and smoothness terms, $\mathcal{G}$ is the 2-dimensional grid defined by the virtual camera, $D_{2\mathrm{D}}(B(\boldsymbol{p}, d_{\boldsymbol{p}}))$ is the data term of the function measured at 3D point $B(\boldsymbol{p}, d_{\boldsymbol{p}})$, obtained by backprojecting image point $\boldsymbol{p}$ at the depth corresponding to label $d_{\boldsymbol{p}}$. $\mathcal{N}_{2\mathrm{D}}$ is the set of interacting nodes defined by a 4-connected neighbourhood in the image grid, and $S_{2\mathrm{D}}(B(\boldsymbol{p}, d_{\boldsymbol{p}}), B(\boldsymbol{q}, d_{\boldsymbol{q}}))$ is the smoothness term, which corresponds to the normal consistency term between 3D points $B(\boldsymbol{p}, d_{\boldsymbol{p}})$ and $B(\boldsymbol{q}, d_{\boldsymbol{q}})$.

In this approach, the data term is computed as:

$$D_{2\mathrm{D}}(\boldsymbol{P}) = \begin{cases} 1 & \text{if } |\mathrm{vis}(\boldsymbol{P})| < n_{\mathrm{vis}}, \\ e^{-\mu \times \frac{\sigma_2(\boldsymbol{P})}{\sigma_3(\boldsymbol{P})}} & \text{otherwise}, \end{cases} \tag{5}$$

where $\mathrm{vis}(\boldsymbol{P})$ indicates the set of reciprocal pairs of cameras from which point $\boldsymbol{P}$ is visible, n_{vis} is a threshold representing the minimum number of reciprocal pairs of cameras for reliable normal estimation, μ is set to $0.2\ln(2)$ to replicate the same weight used in Roubtsova and Guillemaut (2018) and σ_2 and σ_3 are the second and third singular values of W. To prevent any potential numerical stability issues when σ_3 becomes numerically close to zero, a very small ϵ value can be added to the denominator of the exponent in (5).

An important contribution, to enable application to complex 3D scenes, is the introduction of the visibility term. The first criterion to determine visibility is to only consider the cameras whose axes stand at an angle smaller than 80° with respect to the virtual camera axis. Then, occlusions are computed by approximating each point's visibility based on the visibility at its closest point on the surface of the VH. If an intersection is found between the VH and the segment connecting the camera centre to the approximated point, the camera and its reciprocal are considered to be occluded and therefore are not used, as shown in Fig. 3.

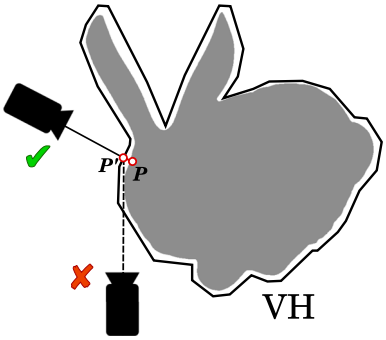

Fig. 3 A point $\boldsymbol{P}$ on the surface is approximated as its closest point $\boldsymbol{P}'$ on the VH before occlusions are taken into consideration for visibility computation

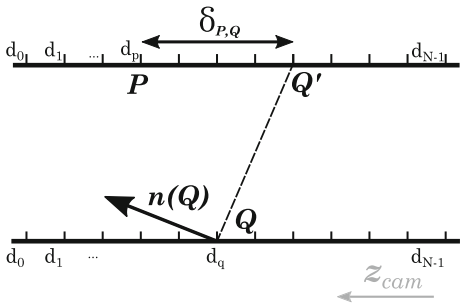

Fig. 4 Illustration of the computation of $\delta_{\boldsymbol{P},\boldsymbol{Q}}$ used to define the smoothness term. Given two neighbouring nodes' depth estimates (d_p and d_q), $\delta_{\boldsymbol{P},\boldsymbol{Q}}$ is computed as the distance between d_p and the estimated surface position at the same node based on d_q and its estimated normal $\boldsymbol{n}(\boldsymbol{Q})$. z_{cam} denotes the unit vector defining the virtual camera axis and oriented such that it is pointing towards the camera

The smoothness function used here is the distance based DNprior (Roubtsova & Guillemaut, 2018), which enforces a smooth surface that is consistent with the normals obtained through HS. This term is calculated as follows:

$$S_{2\mathrm{D}}(\boldsymbol{P}, \boldsymbol{Q}) = \begin{cases} \frac{1}{2}(\delta^2_{\boldsymbol{P},\boldsymbol{Q}} + \delta^2_{\boldsymbol{Q},\boldsymbol{P}}) & \text{if } \delta_{\boldsymbol{P},\boldsymbol{Q}} \text{ and } \delta_{\boldsymbol{Q},\boldsymbol{P}} < t, \\ t^2 & \text{otherwise}, \end{cases} \tag{6}$$

where t is the maximum threshold for $\delta_{\boldsymbol{P},\boldsymbol{Q}}$ and $\delta_{\boldsymbol{Q},\boldsymbol{P}}$. $\delta_{\boldsymbol{P},\boldsymbol{Q}}$ is the distance between point $\boldsymbol{P}$ and the projection of $\boldsymbol{Q}$, perpendicular to its estimated normal, on the pixel where $\boldsymbol{P}$ lies, as illustrated in Fig. 4. It is calculated as follows:

$$\delta_{\boldsymbol{P},\boldsymbol{Q}} = \frac{|\boldsymbol{PQ} \cdot \boldsymbol{n}(\boldsymbol{Q})|}{\boldsymbol{n}(\boldsymbol{Q}) \cdot z_{\mathrm{cam}}}, \tag{7}$$

where $\boldsymbol{PQ}$ is the vector connecting $\boldsymbol{P}$ and $\boldsymbol{Q}$, $\boldsymbol{n}(\boldsymbol{Q})$ indicates the estimated normal at point $\boldsymbol{Q}$ and z_{cam} defines the virtual camera axis. We adopt the convention that z_{cam} is oriented towards the camera. Consequently, $\boldsymbol{n}(\boldsymbol{Q}) \cdot z_{\mathrm{cam}}$ is guaranteed to be positive for any visible surface point. Whenever $\delta_{\boldsymbol{P},\boldsymbol{Q}}$ or

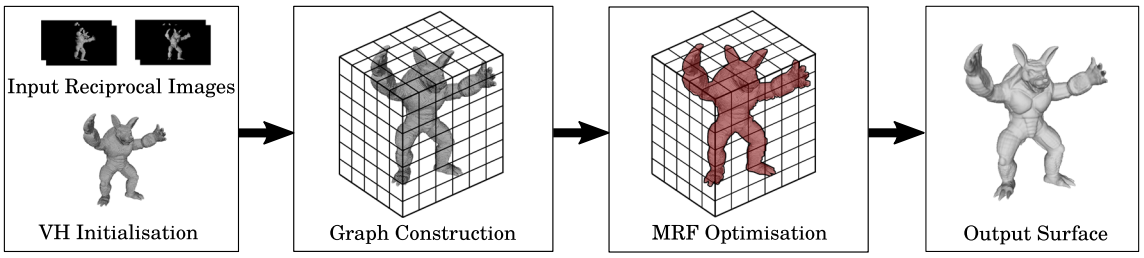

Fig. 5 Pipeline for the volumetric approach to full 3D optimisation. The VH is used for initialisation. Then an orthographic grid is used to compute the probability of voxel occupancy for the surface. The optimisation is then performed using a Markov Random Field (MRF) to obtain the final representation of the object

$\delta_{\boldsymbol{Q},\boldsymbol{P}}$ are greater than a threshold t, dependent on the reconstruction resolution, this term is truncated to t^2 in order to avoid heavy penalties where a strong discontinuity is present on the surface. An illustration of how S_{2D} is computed is shown in Fig. 4.

The energy function is then minimised using Sequential Tree-Reweighted Message Passing (TRW-S) (Kolmogorov, 2015) to obtain the depth maps from each viewing direction. This optimisation framework was chosen because it is able to perform optimisation in a multi-label MRF, while giving strong optimality guarantees. The partial surface reconstructions are then fused together using Poisson surface integration (Kazhdan et al., 2006). This process may lead to inconsistencies across overlapping areas of the partial surfaces, therefore a confidence score is employed to weight the different samples based on their reliability during the fusion process. During the energy function computation of the view-dependent methods, two confidence scores are computed for each point to be used during the surface integration process.

The first confidence score is derived from the saliency and defined from Eq. 5 as follows:

$$C_s(\boldsymbol{P}) = 1 - D_{2D}(\boldsymbol{P}), \tag{8}$$

where $\boldsymbol{P}$ is the selected point after the optimisation process. The second confidence score is derived from the angle between the viewing direction and the estimated normal and is defined as:

$$C_{\boldsymbol{n}}(\boldsymbol{P}) = \boldsymbol{n}(\boldsymbol{P}) \cdot z_{\text{cam}}, \tag{9}$$

where $\boldsymbol{n}(\boldsymbol{P})$ is the estimated normal at 3D point $\boldsymbol{P}$ and z_{cam} is the camera axis of the chosen view, orthographic or perspective depending on the reconstruction method used. The overall confidence score is obtained by multiplying the two confidence scores, that is:

$$C_{s\boldsymbol{n}}(\boldsymbol{P}) = C_s(\boldsymbol{P}) C_{\boldsymbol{n}}(\boldsymbol{P}). \tag{10}$$

Utilising these confidence scores allows to resolve ambiguities for points that are in close proximity, but stem from different views. In particular, the saliency score may appear low in specific camera combinations due to object occlusions or poor visibility, while $C_{\boldsymbol{n}}$ might detect a decline in normal estimation precision at the edges of a reconstructed view or when the surface is heavily slanted with respect to the camera axis. In these scenarios the fusion algorithm may favour the use of a different view with higher confidence scores, to increase the overall reconstruction accuracy.

3.2 Volumetric Full 3D Helmholtz Stereopsis

This second method performs direct optimisation over a voxel grid, where each voxel is labelled as outside the surface, inside, or containing a section of the surface itself. The method enforces coherency between neighbouring voxels on the whole surface, presenting a strong advantage with respect to the fusion of multiple partial 2.5D reconstructions, where the optimisation is performed separately for each partial reconstruction prior to merging. After dividing the volume in a regular voxel grid, a multi-label MRF is constructed, where each node corresponds to a voxel and its labels indicate whether the voxel is occupied by the surface or found inside or outside the object. Weights are then assigned to each voxel depending on their occupancy probability and a regularising term based on the HS estimated normals across neighbouring nodes. The optimisation is performed using a modified version of Iterative Conditional Modes (ICM) (Besag, 1986), where the labelling is changed iteratively to obtain the lowest score possible across local node clusters. The result is finally integrated using Poisson surface reconstruction to obtain the full 3D object representation. See Fig. 5 for an overview of the pipeline.

A 3D orthographic grid, that encompasses the whole object, is first created and each voxel is assigned a node in a multi-label MRF graph. The label set used is the following: $\{I, O, L_0, \ldots, L_{N-1}\}$, where I and O indicate respectively whether the voxel is found inside or outside the reconstructed surface, while the remaining labels are assigned when the sur-

face is crossing inside the voxel. The sampling strategy used consists in subdividing the voxel into equal size subvoxels, where the centre of each subvoxel will correspond to a label between L_0 and L_{N-1}. In this paper, each voxel is subdivided into 27 subvoxels, subdividing by 3 along each dimension. The use of multiple surface labels is employed to generate enough variations in the orientation of the segments connecting neighbouring surface samples, which allows to exploit the normals estimated through HS. Using a single surface label would not allow to fully exploit the normals during the energy function regularisation, as the segments connecting the voxels would only be sampled at 45° steps. It must be noted that sampling the voxel regularly means that the surface may cross a voxel at multiple labels, in which case the algorithm will prioritise the label where a lower weight is achieved for the overall solution.

Performing a full 3D optimisation requires improving visibility estimation with respect to how it was approximated in the view-dependent approaches. To do so, a probabilistic approach is presented here, which is then applied in the methods outlined in this section. A first selection of cameras is performed by approximating the chosen point to their closest neighbour on the initialisation surface and computing occlusions for said point. A further selection is then performed on these cameras by finding the k pairs that have the highest likelihood of producing coplanar w vectors. The parameter k may be chosen to be 3, since this is the minimum number of camera pairs needed to perform HS. The method consists in iteratively selecting all possible combinations of k camera pairs and obtaining their resulting W matrix, computed by stacking the corresponding $\boldsymbol{w}$ from Eq. 3. The subset that satisfies the following equation is then selected:

$$\max_{c_0,\ldots,c_{k-1}} \frac{\sigma_2(W_{c_0,\ldots,c_{k-1}})}{\sigma_3(W_{c_0,\ldots,c_{k-1}})}, \tag{11}$$

where c_0 to c_{k-1} indicate the selected camera pairs and σ_2 and σ_3 are respectively the second and third singular values of the obtained W matrix.

It must be noted that during the visibility computation it is unknown whether the chosen point is actually on the surface or not. This approach maximises the probability of computing the correct visibility for points on the surface. In case of points outside or inside the object, it is expected that the agreement of the camera pairs will generally be low. An added benefit of selecting a fixed number of cameras to perform the HS calculations is that it provides saliency scores that are consistent across the whole surface, making the optimisation process more robust.

Now that the graph construction, visibility handling and label strategy have been established, the energy function used for the MRF optimisation is defined as:

$$E(\boldsymbol{L}) = (1-\beta)\sum_{\boldsymbol{v}\in\mathcal{V}} D_{3\mathrm{D}}(\boldsymbol{v}, L_{\boldsymbol{v}}) + \beta \sum_{(\boldsymbol{v},\boldsymbol{w})\in\mathcal{N}_{3\mathrm{D}}} S_{3\mathrm{D}}(\boldsymbol{v}, L_{\boldsymbol{v}}, \boldsymbol{w}, L_{\boldsymbol{w}}), \tag{12}$$

where $D_{3\mathrm{D}}$ and $S_{3\mathrm{D}}$ are respectively the data and smoothness terms, β is a weight to balance their effects, $\mathcal{V}$ indicates the volume in which the 3D grid is constructed and $\mathcal{N}_{3\mathrm{D}}$ is the set of interacting nodes defined by a 6-connected neighbourhood in the 3D grid.

The data term is based on the HS saliency as in the methods previously presented. The higher the ratio between the singular values σ_2 and σ_3 of matrix W, the higher the probability of the point being located on the surface. In addition to the previous methods, however, the additional inside and outside labels are considered in this approach and the full data term calculation is performed as follows:

$$D_{3\mathrm{D}}(\boldsymbol{v}, L_{\boldsymbol{v}}) = \begin{cases} 0 & \text{if } L_{\boldsymbol{v}} \in \{I, O\}, \\ 1 & \text{if } L_{\boldsymbol{v}} \notin \{I, O\} \text{ and} \\ & |\mathrm{vis}(M(\boldsymbol{v}, L_{\boldsymbol{v}}))| < n_{\mathrm{vis}}, \\ e^{-\mu \times \frac{\sigma_2(M(\boldsymbol{v},L_{\boldsymbol{v}}))}{\sigma_3(M(\boldsymbol{v},L_{\boldsymbol{v}}))}} & \text{otherwise,} \end{cases} \tag{13}$$

where $M(\boldsymbol{v}, L_{\boldsymbol{v}})$ indicates the 3D position of the surface point at node $\boldsymbol{v}$ when assigned label $L_{\boldsymbol{v}}$ and $\mathrm{vis}(\boldsymbol{P})$ represents the set of camera pairs from which point $\boldsymbol{P}$ is visible. Usually n_{vis} is set to 5, which is greater than the minimum number of camera pairs required in HS to avoid geometric ambiguities. Whenever a voxel is assigned a surface label $(L_0, \ldots, L_{N-1})$ and is visible in a sufficient number of cameras, the weight assigned is computed as the data term in the view-dependent methods as shown in Eq. 5.

The smoothness term is used as a regularising constraint to ensure that neighbouring voxels have coherent normals and that there is always a surface voxel between an outside and inside voxel pair, avoiding holes in the final surface. It is computed as follows:

$$S_{3\mathrm{D}}(\boldsymbol{v}, L_{\boldsymbol{v}}, \boldsymbol{w}, L_{\boldsymbol{w}}) = \begin{cases} \Gamma(M(\boldsymbol{v}, L_{\boldsymbol{v}}), M(\boldsymbol{w}, L_{\boldsymbol{w}})) & \text{if } L_{\boldsymbol{v}}, L_{\boldsymbol{w}} \in \\ & \{L_0, \ldots, L_{N-1}\}, \\ \infty & \text{if } L_{\boldsymbol{v}}, L_{\boldsymbol{w}} \in \{I, O\} \\ & \text{and } L_{\boldsymbol{v}} \neq L_{\boldsymbol{w}}, \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

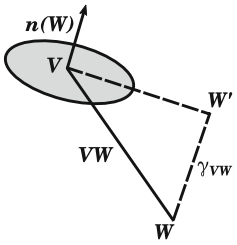

Fig. 6 Illustration of the computation of $\gamma(\boldsymbol{V}, \boldsymbol{W})$. The term is based on the assigned label of node $\boldsymbol{W}$ and its predicted position based on the label of $\boldsymbol{V}$ and its estimated normal $\boldsymbol{n}(\boldsymbol{W})$

where

$$\Gamma(\boldsymbol{V}, \boldsymbol{W}) = \begin{cases} \frac{1}{2}(\gamma^2_{\boldsymbol{V},\boldsymbol{W}} + \gamma^2_{\boldsymbol{W},\boldsymbol{V}}) & \text{if } \gamma_{\boldsymbol{V},\boldsymbol{W}} \text{ and } \gamma_{\boldsymbol{W},\boldsymbol{V}} < t, \\ t^2 & \text{otherwise,} \end{cases} \quad (15)$$

represents the normal consistency between points $\boldsymbol{V}$ and $\boldsymbol{W}$. t is used as a truncation term for $\gamma_{\boldsymbol{V},\boldsymbol{W}}$ which is defined as follows:

$$\gamma_{\boldsymbol{V},\boldsymbol{W}} = |\boldsymbol{VW} \cdot \boldsymbol{n}(\boldsymbol{W})|, \quad (16)$$

where $\boldsymbol{VW}$ is the vector connecting points $\boldsymbol{V}$ and $\boldsymbol{W}$, while $\boldsymbol{n}(\boldsymbol{W})$ indicates the unit normal estimated via HS at point $\boldsymbol{W}$. This term represents the distance between $\boldsymbol{W}$ and the plane perpendicular to $\boldsymbol{n}(\boldsymbol{V})$ intersecting point $\boldsymbol{V}$. Figure 6 illustrates how this term is calculated. The ∞ term is used to constrain inside and outside voxels to be separated by surface voxels, thus avoiding an empty solution where all nodes are either labelled to be inside or outside. The truncation is performed on the regularisation term to avoid heavy penalties where a corner may result in two neighbouring voxels having severely different normals.

Once the graph has been initialised, optimisation is performed using a tailored version of ICM (Besag, 1986). ICM is an exhaustive search algorithm that iterates through an MRF graph and changes one variable at a time, by trying to optimise its local neighbourhood cost. In its classic formulation, ICM would not work in this scenario because of the constraint on the surface. Namely, changing the label of a surface node to be either outside or inside would result in a hole on the surface, which is currently prevented by having an infinite weight when outside and inside voxels are neighbours. However, by changing two neighbouring variables at a time and considering all surrounding nodes to compute the cost change, the surface can be shifted closer to its optimal solution through multiple iterations. Only tuples where one node is on the current surface of the reconstruction are considered at each iteration. To compute their local neighbourhood cost, all possible configurations of said tuple and their neighbours are considered, selecting the solution with the lowest energy. If the problem is initialised close to the actual surface, this step typically converges after a small number of iterations. During experimentation it was attempted to use TRW-S to perform the optimisation, however, due to the high number of nodes and low number of labels, TRW-S obtained poor results, while ICM proved to be suitable in this specific scenario.

Finally, the nodes labelled to be on the surface are extracted together with their Helmholtz estimated normals and integrated using Poisson surface reconstruction to obtain a mesh representation.

3.3 Mesh-Based Full 3D Helmholtz Stereopsis

The key idea of this third approach is to introduce a single-step method to perform full 3D reconstruction from a coarse initialisation, providing a mechanism to perform global optimisation and recover a solution with strong optimality properties through the use of state-of-the-art MRF solvers. This method offers some major advantages with respect to the volumetric approach, which is much more computationally expensive, by providing a mesh-based optimisation that seeks to find the optimal 3D positions of the vertices in the solution space. The target surface is obtained through a Maximum a Posteriori (MAP) approach using an MRF graph.

The 3D reconstruction is performed through a pipeline that is illustrated in Fig. 7. To initialise the method and identify the search space over which the optimisation is performed, two surfaces are defined. The first one corresponds to the outer boundary of the solution space and must completely encompass the object. For instance, the VH of the object or an accordingly dilated approximation of a previous solution, obtained from a different technique, could be used. The second surface must, instead, be completely inside the object, while maintaining a similar topology to the outer surface; this can be achieved by carving the outer surface. Non-rigid Iterative Closest Point (ICP) (Audenaert et al., 2019) is used to draw correspondences between the two surfaces, allowing to match key features between the surfaces despite their difference in scale. The result is a dense matching between the two surfaces, where a point of the target surface is necessarily found between each pair of registered vertices.

These correspondences are used to construct a multi-label MRF graph. Each pair of corresponding vertices between the two surfaces will be assigned to a node and its neighbours will be established depending on the surface topology. In practice, the surface is represented as a 3D mesh from which each edge corresponds to a graph edge, defining the neighbourhood of the two bound vertices. Each node is then assigned a set of labels $\{L_0, L_1, \ldots, L_{N-1}\}$, where each label indicates a 3D point on the segment connecting the two surfaces at the node's corresponding vertices. In particular, L_0 will coincide with the vertex on the outer surface, L_{N-1} with the corresponding vertex on the inner surface and all the intermediate labels will be spaced regularly in between them. In the remainder of this section, the energy function used to perform the optimisation is detailed.

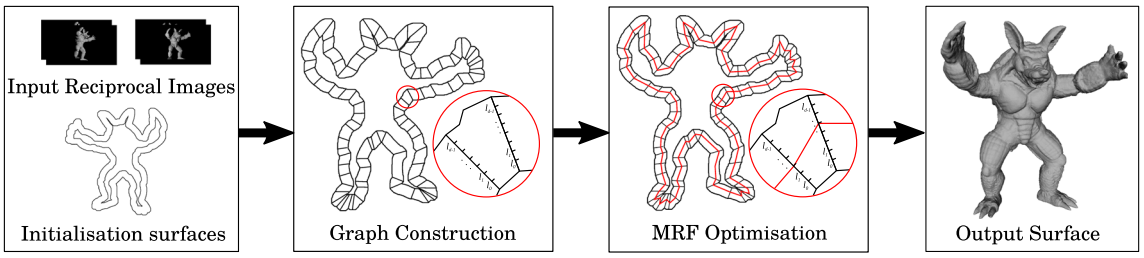

Fig. 7 Pipeline overview for the mesh-based approach. Two initialisation surfaces that contain the object's boundary are used to construct an MRF graph where each node corresponds to a pair of points on the surfaces and their labels are the points between them. The output surface is obtained by performing global optimisation on the graph and connecting the nodes estimated positions

To reconstruct the target surface, an energy function is defined as follows:

$$E(\boldsymbol{L}) = (1-\alpha)\sum_{v\in\mathcal{M}} D_{\mathrm{M}}(X(\boldsymbol{v}, L_{\boldsymbol{v}})) + \alpha \sum_{(\boldsymbol{v},\boldsymbol{w})\in\mathcal{N}_{\mathrm{M}}} S_{\mathrm{M}}(X(\boldsymbol{v}, L_{\boldsymbol{v}}), X(\boldsymbol{w}, L_{\boldsymbol{w}})), \tag{17}$$

where $\boldsymbol{L}$ indicates the labels assigned across the entire set of nodes, α is a weighting factor to balance the effect of data and smoothness terms and $\mathcal{N}_{\mathrm{M}}$ is the set of interacting nodes defined by the connectivity of the mesh representing the optimised surface $\mathcal{M}$. D_{M} and S_{M} are respectively the data and smoothness terms, calculated for all the nodes and edges of the graph, while the operator $X(\boldsymbol{v}, L_{\boldsymbol{v}})$ is used to identify the resulting position when node $\boldsymbol{v}$ is assigned the label $L_{\boldsymbol{v}}$.

The data term is based on the HS saliency measure and computed in a similar fashion to Eq. 5 whenever the point is deemed visible by at least a certain number of cameras:

$$D_{\mathrm{M}}(\boldsymbol{P}) = \begin{cases} 1 & \text{if } |\mathrm{vis}(\boldsymbol{P})| < n_{\mathrm{vis}}, \\ e^{-\mu \times \frac{\sigma_2(\boldsymbol{P})}{\sigma_3(\boldsymbol{P})}} & \text{otherwise}, \end{cases} \tag{18}$$

where σ_2 and σ_3 indicate the second and third singular values of the W matrix, and $\mathrm{vis}(\boldsymbol{P})$ is the set of camera pairs from which point $\boldsymbol{P}$ is visible. Points that are not visible from enough cameras are given a strong weight that still allows for points affected by self-occlusions to be reconstructed, in accordance with the neighbouring points which may be visible.

The smoothness term serves as a regularising weight to ensure the surface is smooth and consistent with the photometric normals calculated through HS. It is based on a depth disparity measure, here referred to as $\delta(\boldsymbol{V}, \boldsymbol{W})$, calculated between pairs of neighbouring nodes. $\delta(\boldsymbol{V}, \boldsymbol{W})$ represents the distance between a point and its predicted position based on the estimated normal of its neighbour. It is calculated as follows:

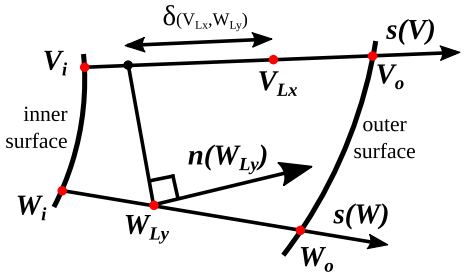

Fig. 8 Illustration of the computation of the smoothness term. The point $\boldsymbol{W}_{L_y}$ is projected perpendicularly to its estimated normal towards the segment $\boldsymbol{V}_i\boldsymbol{V}_o$ and the error is measured as the distance between the projection and point $\boldsymbol{V}_{L_x}$

$$\delta(\boldsymbol{V}, \boldsymbol{W}) = \frac{\boldsymbol{VW}\cdot\boldsymbol{n}(\boldsymbol{W})}{\boldsymbol{n}(\boldsymbol{W})\cdot\boldsymbol{s}(\boldsymbol{V})}, \tag{19}$$

where $\boldsymbol{VW}$ indicates the vector connecting the two points and $\boldsymbol{s}(\boldsymbol{V})$ is a unit vector representing the direction of the segment that connects inner and outer surfaces at the node corresponding to point $\boldsymbol{V}$. The disparity error is computed as the difference between point $\boldsymbol{V}$ and the projection of $\boldsymbol{W}$ perpendicular to its estimated normal $\boldsymbol{n}(\boldsymbol{W})$ towards said segment. Figure 8 illustrates the definition of the smoothness term.

This term is a generalisation to a perspective sampling in full 3D of the depth disparity measure presented in Eq. 7. Moreover, the error measure presented here is more discriminative than the one used in Eq. 14, which is tied to the voxel size chosen and where strong discontinuities do not result in a considerable error. In contrast, the proposed distance penalises more heavily depth and normal assignments which are inconsistent between neighbouring nodes, which are not bound by the volume sampling resolution.

The smoothness term is then computed as the average of the squared disparity terms for the two neighbours and it is

truncated at a threshold of t^2:

$$S_{\mathrm{M}}(\boldsymbol{V}, \boldsymbol{W}) = \begin{cases} \frac{1}{2}(\delta(\boldsymbol{V}, \boldsymbol{W})^2 + \delta(\boldsymbol{W}, \boldsymbol{V})^2) & \text{if } \delta(\boldsymbol{V}, \boldsymbol{W}) \text{ and} \\ & \delta(\boldsymbol{W}, \boldsymbol{V}) < t, \\ t^2 & \text{otherwise.} \end{cases} \tag{20}$$

The threshold is used to allow for natural discontinuities and it is also used where occlusions do not allow HS to produce an estimated normal.

The final aspect taken into consideration in this methodology is the technique that can be used to perform the final optimisation. The energy function chosen to represent the problem violates the submodularity constraint and non-submodular functions cannot be properly minimised by classic graph-cut approaches, as indicated in Kolmogorov and Rother (2007). However, many techniques exist to approximate the solution of a non-submodular function with a high degree of confidence. The approach used here is Tree-Reweighted Message Passing (TRW) (Wainwright et al., 2005) in its more recent formulation called TRW-S (Kolmogorov, 2006, 2015), which, contrary to TRW, guarantees that the energy lower bound does not decrease during optimisation and introduces the condition of weak tree agreement to identify local maxima in the energy bound.

After the final solution is obtained, it is not necessary to perform surface integration, since the initialisation surface topology is maintained in the obtained result. It was however found beneficial to perform some integration using Poisson surface reconstruction on the obtained vertices, as an extra regularisation step to leverage the normals obtained using HS. The extra integration step is also able to smooth some artefacts caused by minor differences which may exist between the topology of the outer and inner surfaces.

4 Experimental Evaluation

This section performs a comparative evaluation of the proposed approaches using both synthetic and real datasets. The methods evaluated will be referred to thereafter as: VH for SfS; VDo for orthographic fused view-dependent HS (Sect. 3.1); VDp for perspective fused view-dependent HS (Sect. 3.1); 3DHSv for volumetric full 3D HS (Sect. 3.2) and 3DHSm for mesh-based full 3D HS (Sect. 3.3). An additional method referred to as 3DHSc is also introduced. This uses a volumetric graph-cut approach (Vogiatzis et al., 2007) which we adapted to HS for comparison purposes. The method estimates the voxel occupancy in a grid by performing graph-cut on a binary-label MRF. The data term is a constant ballooning term which is applied to outside voxels to avoid an empty solution. The smoothness term is computed at the edges of neighbouring voxels and is based on the HS saliency measure, similarly to the data terms proposed in the previous methods. The proposed approaches are also compared against state-of-the-art MVS, MVPS and neural radiance field-based techniques in the case of real scenes. An ablation study analysing the effects of the proposed confidence scores on the performance of the view-dependent approaches (VDo and VDp) is also included in Appendix A.

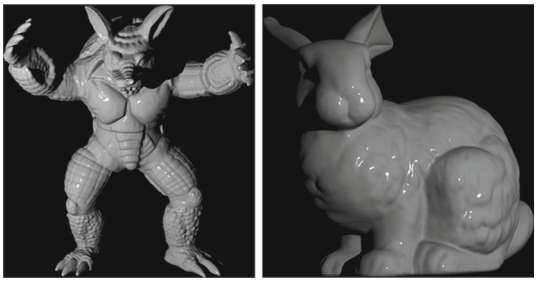

Fig. 9 Examples images from the *Armadillo* and *Bunny* datasets

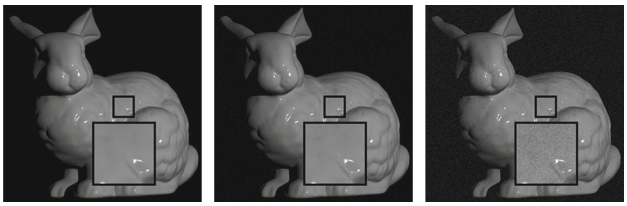

Fig. 10 Illustration of the different image noise levels applied to the synthetic dataset: no noise (left), Gaussian noise with standard deviation 0.01% (middle) and 0.1% (right). The inset images show magnifications of the portion inside the square

4.1 Evaluation on Synthetic Scenes

A novel synthetic dataset was generated using the Stanford Bunny (Turk & Levoy, 1994) and the Armadillo (Krishnamurthy & Levoy, 1996). To ensure the synthesised images are physically plausible, and in particular satisfy Helmholtz reciprocity, images were rendered using the modified Phong reflectance model (Lewis, 1994). The model combines a diffuse and a specular component, with the BRDF defined as:

$$f(k_d, k_s) = k_d \frac{1}{\pi} + k_s \frac{\frac{1}{r} + 2}{2\pi} (\boldsymbol{h} \cdot \boldsymbol{n})^{\frac{1}{r}}, \tag{21}$$

where k_s and k_d respectively represent the diffuse and specular coefficients. The specular component further depends on the surface roughness r and the angle between the normal $\boldsymbol{n}$ at the observed surface point and the vector $\boldsymbol{h}$ bisecting the incoming and outgoing light directions. In each case, 40 reciprocal pairs of images were generated by sampling viewpoints on a sphere of radius 600 mm with the object located at the centre of the sphere. *Armadillo* and *Bunny* are 151 mm and 153 mm tall respectively. The camera used to render the

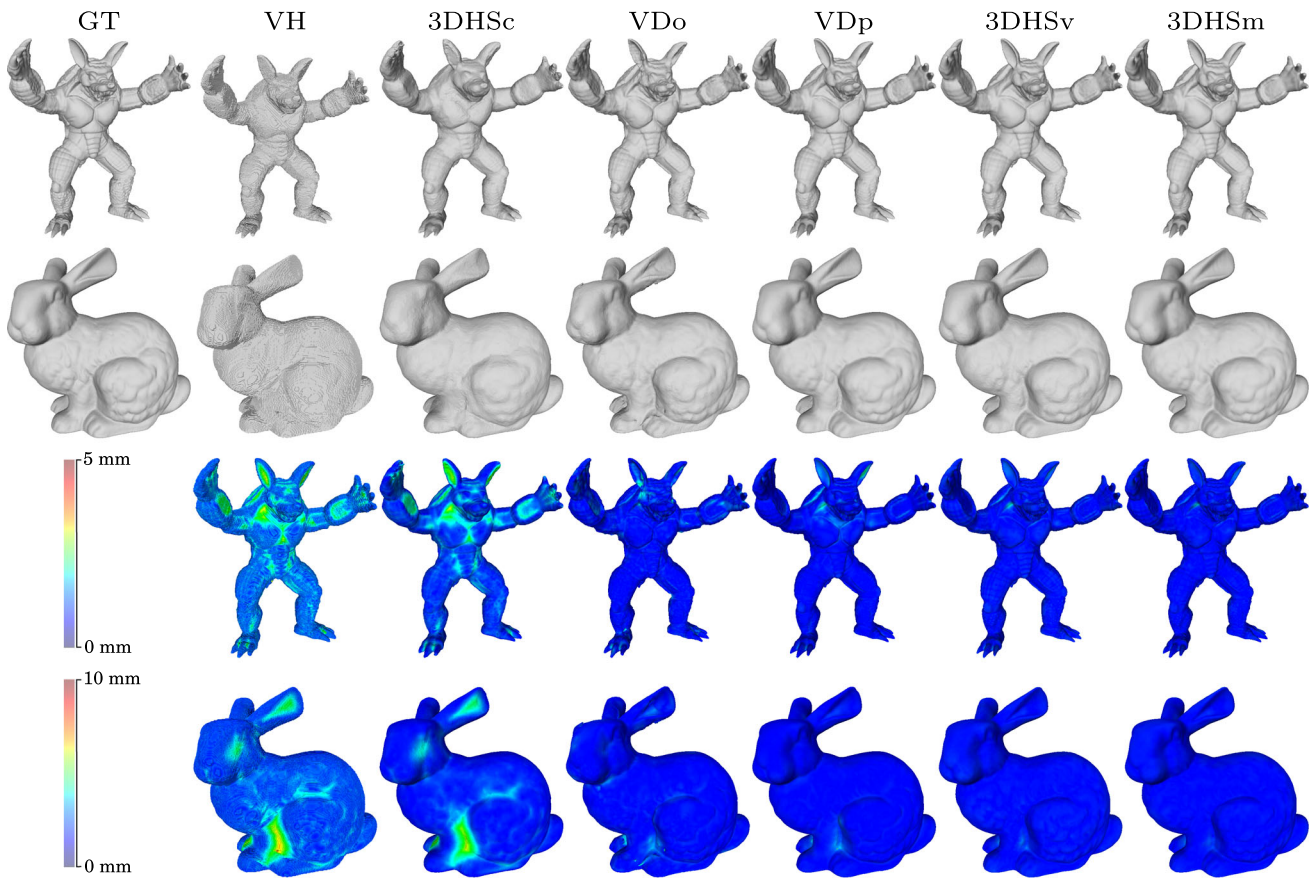

Fig. 11 Results including error maps obtained on the synthetic scenes *Armadillo* and *Bunny* with 0.01% noise level

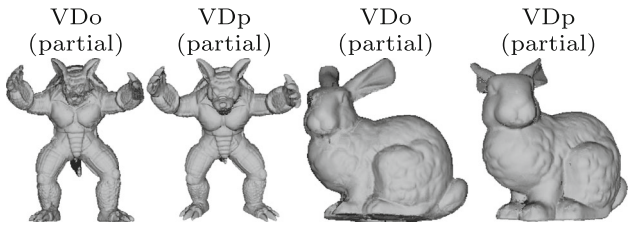

Fig. 12 Examples of partial reconstructions for the *Armadillo* and *Bunny* datasets produced using the orthographic and perspective view-dependent approaches. It must be noted that the reconstructed portions differ slightly between the two methods due to the different viewpoints and image formation models they use

views had a horizontal field of view of 40°, equivalent to a focal length of 2638 pixels/mm. Images were rendered at a resolution of 1920 × 1080. Different datasets were generated by adding Gaussian noise at three different levels with standard deviation of 0%, 0.01% and 0.1% of the full 16-bit image range. Example images in case of the 0.01% noise level are shown in Fig. 9 for each object. Figure 10 shows a close-up of an image corrupted at the different noise levels in the case of the *Bunny* dataset. Ground truth models are shown in the first column of Fig. 11.

Figure 11 shows the reconstructions for the different methods and their respective error maps in the case of the intermediate noise level (with standard deviation 0.01%). Figure 12 also shows an example of intermediate view-dependent reconstruction for each object in the case of the VDo and VDp methods. As can be observed, the VH method performs poorly, especially with regards to reconstructing concavities. In comparison, 3DHSc achieves better results due to the use of HS for normal estimation, but results still lack surface detail. This shows the limitations of relying solely on HS saliency and a ballooning term for energy regularisation. VDo achieves better results by introducing a tailored regularisation term enforcing the consistency of the depth and normal estimates. However, artefacts are present on the surface due to the fusion of the separately computed partial surfaces, which present some imperfections due the use of orthographic cameras. While this is somewhat mitigated by the use of Poisson surface reconstruction and the confidence scores, some fine details present on the surfaces are lost in the process. VDp obtains notably better results, presenting considerably fewer artefacts and achieving a faithful reconstruction of the object. Improvements resulting from switching to a perspective grid with VDp are clearly visible in

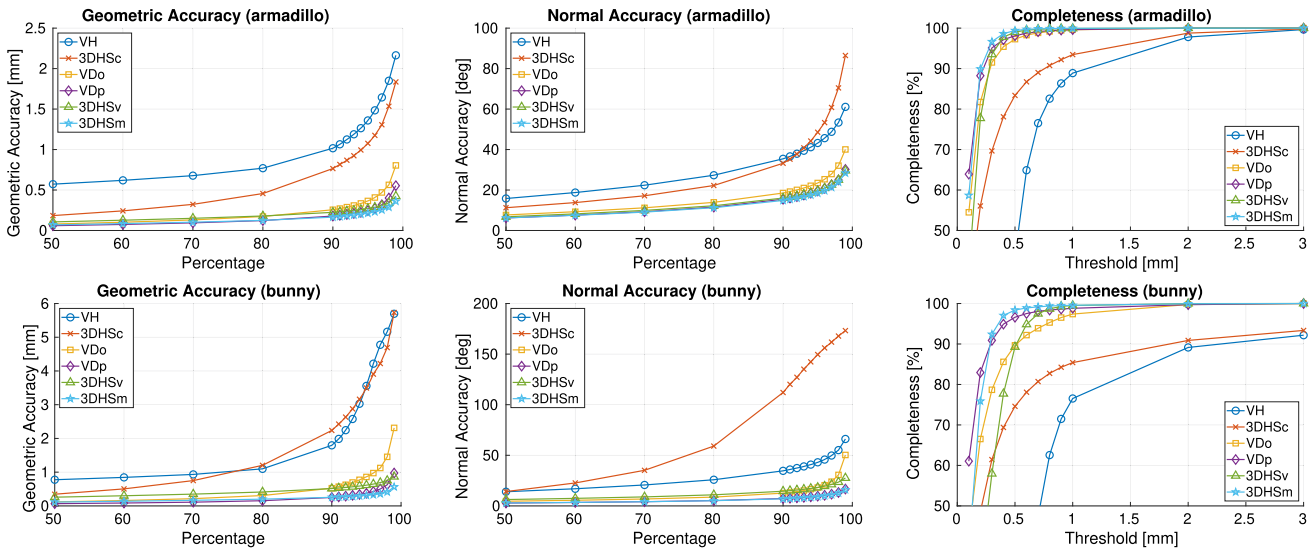

Fig. 13 Geometric accuracy, normal accuracy and completeness graphs for the synthetic scenes *Armadillo* and *Bunny*

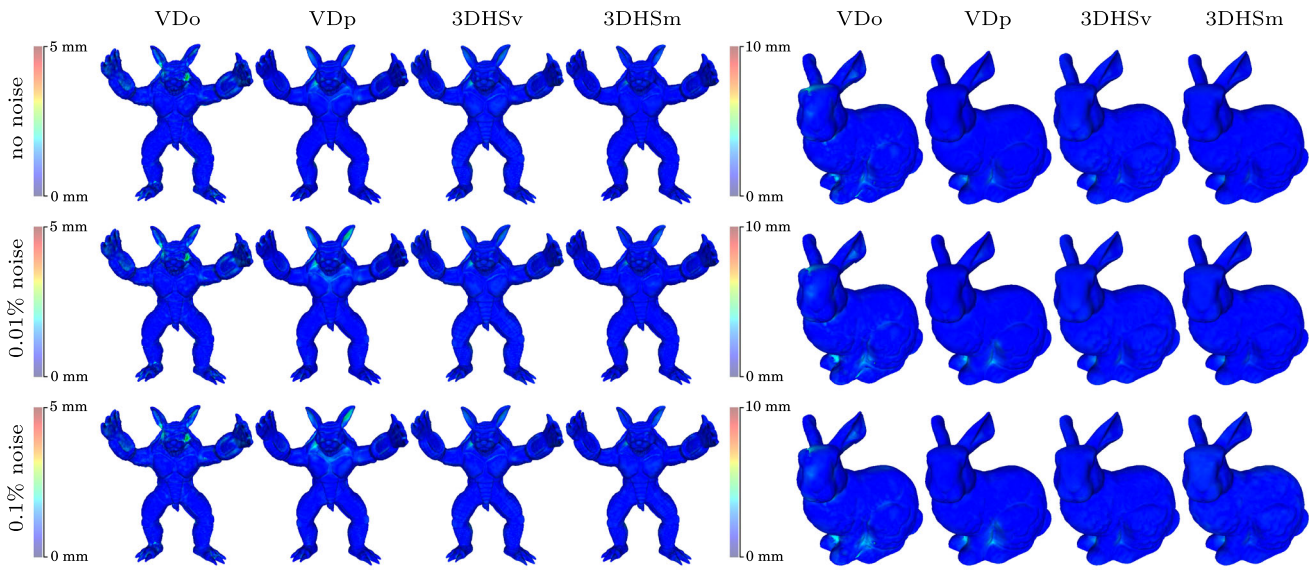

Fig. 14 Error maps of the results obtained on the *Armadillo* and *Bunny* datasets at different levels of noise for the proposed methods

the partial reconstructions shown in Fig. 12, which ultimately results in better final reconstructions after fusion into complete models, compared to their orthographic counterparts obtained using VDo. Results obtained using 3DHSv overcome some of the limitations of the fused view-dependent approaches, by introducing a volumetric optimisation step. However, the use of ICM during optimisation is iterative and unable to fully retrieve some minute concavities present on the surfaces. Furthermore, the implementation of the method as a voxel grid ties its results to the resolution chosen, which is hindered by the computational resources available since memory consumption grows cubically with resolution in the case of a vanilla voxel implementation. Experiments reported here were run on a server with 60GB memory. These limitations could be alleviated through use of an octree implementation, however this was not considered in our implementation. In contrast, 3DHSm is able to retrieve fine surface details such as the shell grooves of the *Armadillo* and the facial features of the *Bunny*.

A quantitative evaluation is performed by computing the Middlebury geometric accuracy, normal accuracy and completeness scores (Seitz et al., 2006) for all methods. These are shown in Fig. 13. The graphs show how the geometric and normal accuracies vary for each method as a wider percentage of the surface is taken into consideration (shown on the horizontal axis), while for the completeness a varying dis-

Table 1 Results obtained on the synthetic scenes using all methods under different noise levels

	Method	No Noise			0.01% Noise			0.1% Noise		
		G. Acc. ↓	N. Acc. ↓	Comp. ↑	G. Acc. ↓	N. Acc. ↓	Comp. ↑	G. Acc. ↓	N. Acc. ↓	Comp. ↑
Armadillo	VH	1.02	35.41	43.6	1.02	35.41	43.6	1.02	35.41	43.6
	3DHSc	0.76	32.69	83.7	0.77	33.21	83.4	0.75	32.73	83.0
	VDo	0.26	18.61	97.3	0.26	18.51	97.3	0.26	18.57	97.1
	VDp	0.17	15.01	98.7	0.18	15.44	98.1	0.19	15.49	98.1
	3DHSv	0.22	16.02	99.0	0.22	16.03	99.0	0.22	16.00	99.0
	3DHSm	**0.16**	**14.71**	**99.5**	**0.16**	**14.84**	**99.4**	**0.18**	**14.97**	**99.3**
Bunny	VH	1.8	34.62	13.9	1.80	34.62	13.9	1.80	34.62	13.9
	3DHSc	2.51	93.95	75.5	2.24	112.06	74.6	2.30	106.16	74.1
	VDo	0.49	13.22	91.2	0.52	12.59	89.7	0.57	12.56	88.3
	VDp	**0.22**	7.02	97.6	**0.25**	7.28	96.6	**0.27**	**7.56**	96.1
	3DHSv	0.30	8.46	97.6	0.51	14.47	89.3	0.29	8.75	**97.6**
	3DHSm	0.23	**6.80**	**98.6**	**0.25**	**6.96**	**98.4**	0.38	7.64	96.8

Geometric accuracy (expressed in mm and denoted by G. Acc.) and normal accuracy (expressed in degrees and denoted by N. Acc.) are computed at a 90% threshold, while the completeness (expressed in % and denoted by Comp.) is obtained at a threshold of 0.5 mm. Top performers are marked in bold

Fig. 15 Acquisition setup consisting of a pair of cameras fitted with external flashes mounted on each camera lens

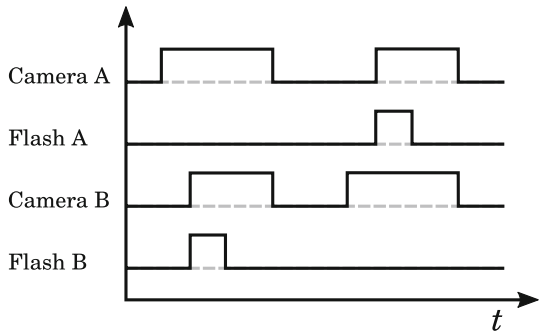

Fig. 16 Timing graph showing how the cameras synchronisation is handled between the camera pairs to capture both HS image and direct flash image with consistent exposure

tance threshold is considered. The graphs confirm that VH and 3DHSc produce poor results, highlighting the impossibility of reconstructing concavities in the first approach and the lack of an effective regularising term in the second method. It can then be observed that VDo, VDp, 3DHSv and 3DHSm significantly outperform the other approaches, achieving significantly better geometric accuracies, normal accuracies and completeness on both objects. VDo performs slightly worse on the *Bunny* dataset, but overall still obtains submillimetre accuracy and an above 90% completeness on thresholds below a millimetre. It can be noted that VDp achieves better results than 3DHSv, despite being based on fusing multiple partial reconstructions. This is likely due to the limitation imposed by the resolution used in the 3DHSv experiments as mentioned previously. In contrast, 3DHSm obtains the best results thanks to the global optimisation performed directly on the mesh vertices.

4.2 Analysis of Robustness to Image Noise

To assess robustness, the different approaches are evaluated on the datasets generated with three different noise levels with standard deviation of 0%, 0.01% and 0.1% of the full 16 bit image range. Figure 14 shows the error maps on the fully reconstructed objects with respect to the ground truth at those different noise levels. Table 1 shows the geometric accuracy, normal accuracy and completeness for each noise level for the typical 90% accuracy and 0.5 mm completeness thresholds. Overall only a small deterioration can be noticed in the results when affected by noise. Areas such as the shell, ears and neck area of the *Armadillo* and the concavities between the legs of the *Bunny* show the main differences between the presented approaches. As it can be observed, in each method the errors are not significantly increased by the introduction of noise, which demonstrates the robustness of the different approaches proposed. This can be explained by the use of

Fig. 17 Example images for the *Fox*, *Corgi*, *Llama*, *Bee*, *Duck* and *Giraffe* datasets

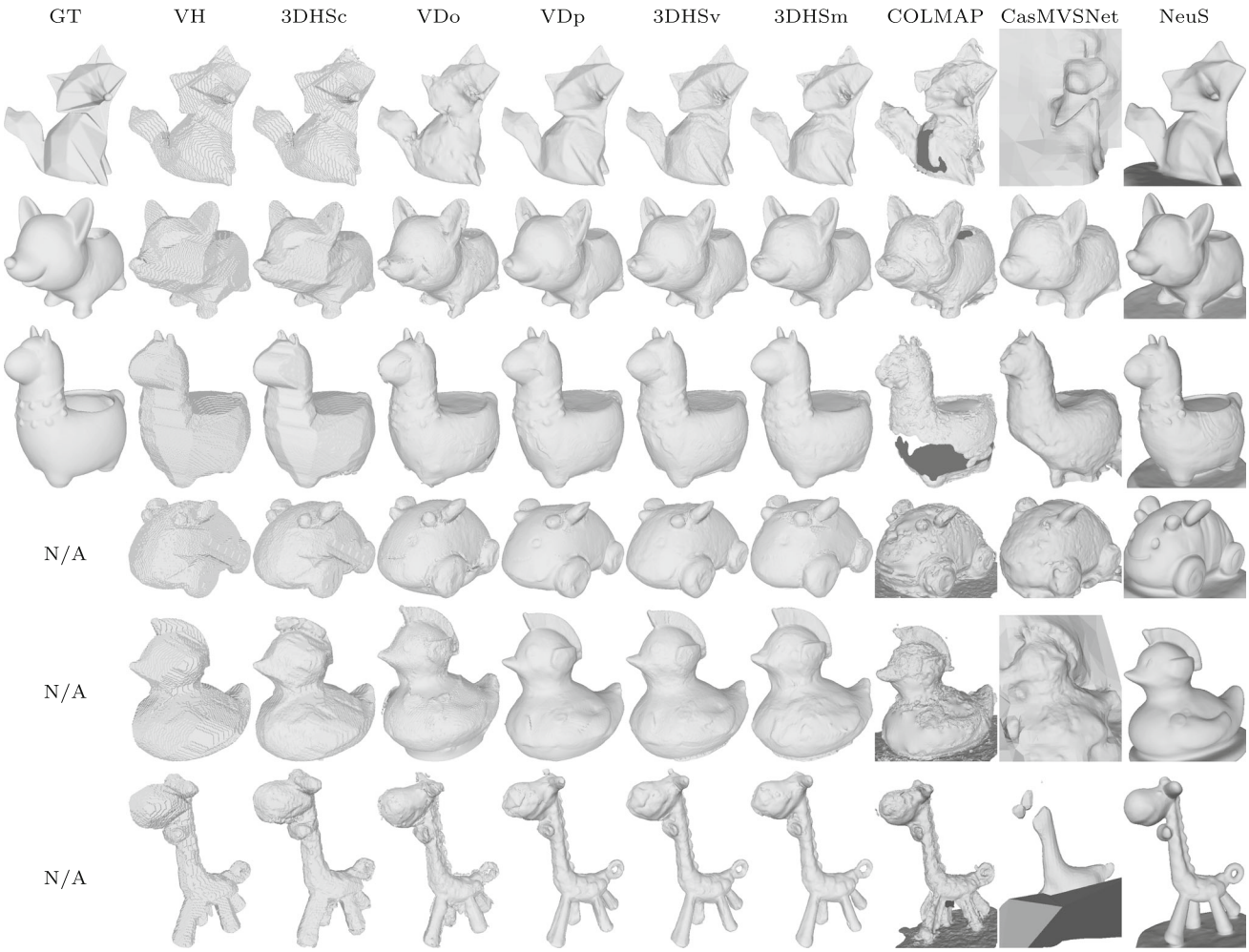

Fig. 18 Results obtained on the real datasets *Fox*, *Corgi*, *Llama*, *Bee*, *Duck* and *Giraffe* for the different methods

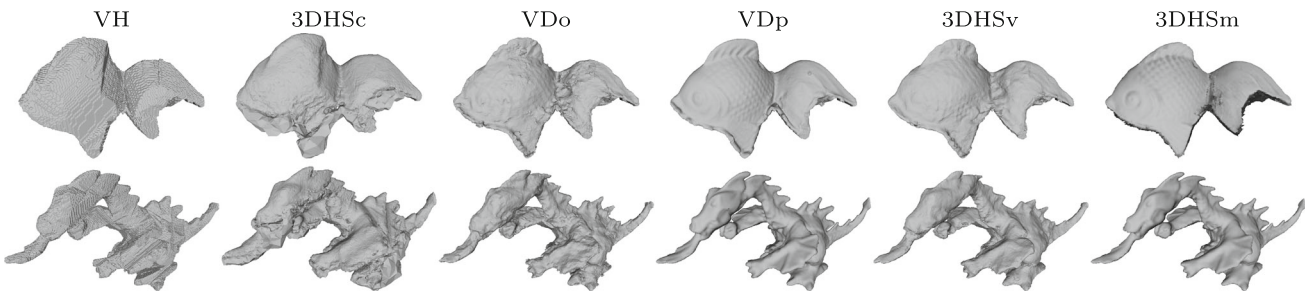

Fig. 19 Results obtained on the real datasets *Fish* and *Dragon* for the different methods

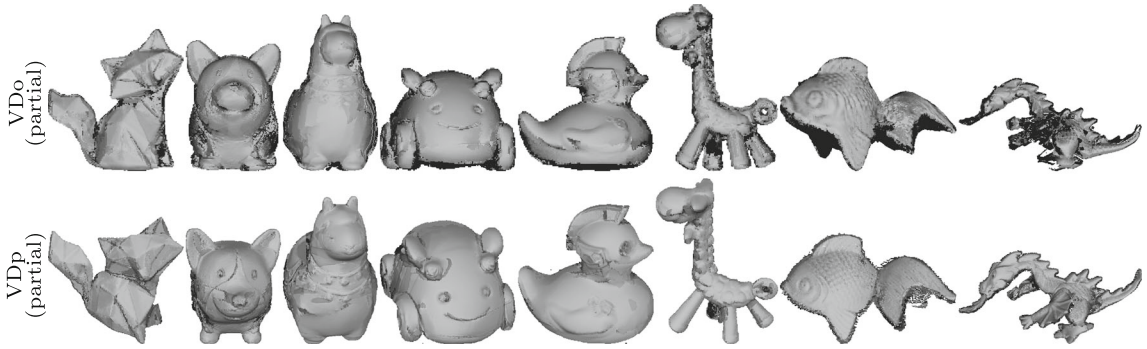

Fig. 20 Examples of partial reconstructions obtained using the orthographic and perspective view-dependent methods on the *Fox*, *Corgi*, *Llama*, *Bee*, *Duck*, *Giraffe*, *Fish* and *Dragon* datasets. Reconstructed portions differ slightly between the approaches due to different viewpoints and image formation models

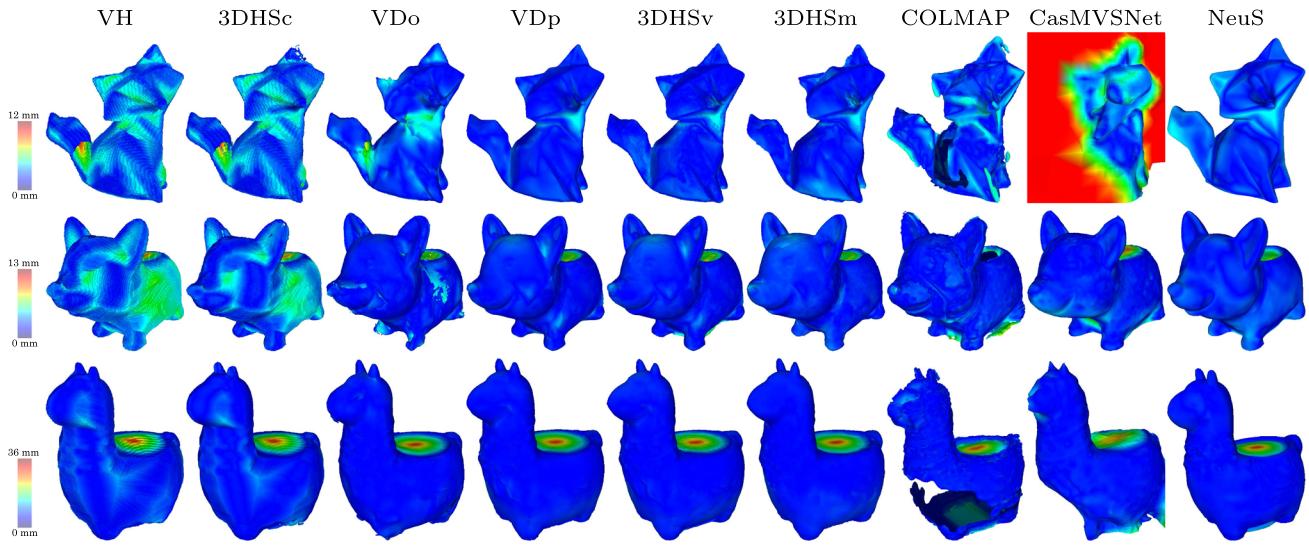

Fig. 21 Error maps for the results obtained on the real datasets *Fox*, *Corgi* and *Llama* for the different methods

window-based averaging to compute the data terms and also the regularisation approaches that both mitigate the effects of noise.

4.3 Evaluation on Real Scenes

Two real datasets are used for evaluation: a novel dataset released with this paper and the *Dragon* and *Fish* scenes from Delaunoy et al. (2010). The new dataset was acquired using a pair of Canon EOS 5DS cameras equipped with Canon Macro Ring Lite MR-14 EX II flashes, which are external macro flashes mounted around each lens to approximate camera/light collocation. The flash on a given lens is synchronised with the other camera to allow acquisition of a reciprocal pair by triggering each camera separately. Moreover, synchronisation of the two cameras allows acquisition of an extra pair of images in which the object is lit directly by the flash mounted on the same camera capturing the image, which can be used for segmentation. Figures 15 and 16 show respectively the acquisition setup and the timing of the left and right cameras and flashes during the acquisition of a reciprocal pair. This provides a flexible setup for HS capture which can either be static with the object placed on a turntable or moved around the object. The former configuration was used to acquire this dataset. The dataset is acquired at a resolution of 2928×4368 and comprises six objects: *Fox*, *Corgi*, *Llama*, *Bee*, *Duck* and *Giraffe* (see Fig. 17). These present several types of materials with varying BRDFs and are for the most part untextured, making them difficult to reconstruct with traditional methods and without prior knowledge over their reflectance properties. *Duck* also presents some subsurface scattering. For each object, 20 image pairs were captured. The cameras were positioned at about 1.5 m from the object and each pair had a baseline of about 60 cm. To perform the capture, the objects were placed on a rotating table, with each image pair taken after a rotation of 18° from the previous pair.

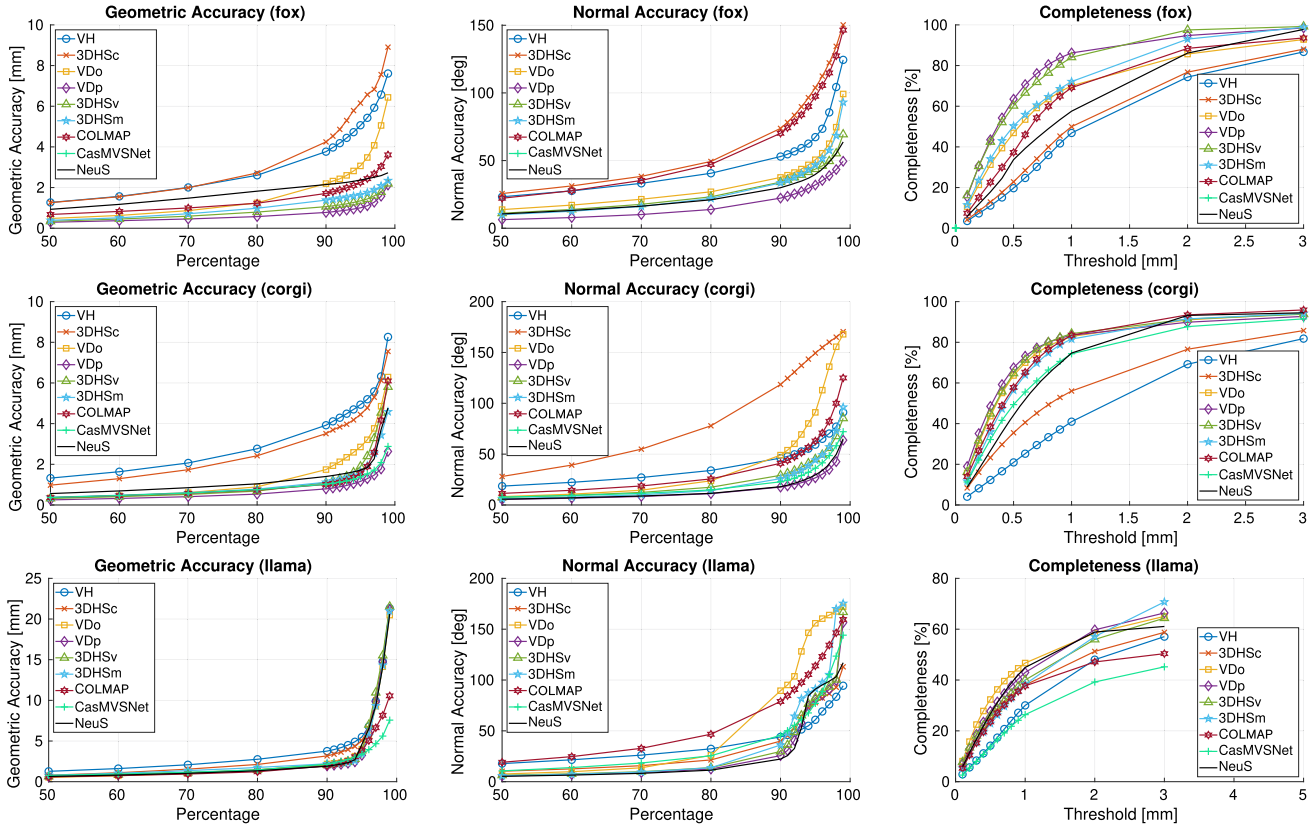

Fig. 22 Geometric accuracy, normal accuracy and completeness graphs for the real datasets *Fox*, *Corgi* and *Llama* for the different HS methods

Figures 18 and 19 show the results obtained on the real scenes for each approach. Further, Fig. 20 shows examples of partial reconstructions for the view-dependent methods. As can be observed, the VH reconstructions are coarse. The 3DHSc method also produces low quality results due to the poor regularisation it relies on. The VDo results are somewhat noisy due to the fusion of multiple 2.5D surfaces that are sometimes incoherent. This can be observed especially at the base of the tail of the *Fox* and *Fish* or in the nose area of the *Corgi*. The perspective approach VDp significantly improves the results with respect to the orthographic method, thanks to the possibility of leveraging segmentation information during the reconstruction. The use of perspective cameras has the advantage of reconstructing views that have been captured directly in the input images, reducing possible effects from occlusions, which can be observed in the *Dragon* and *Fish* datasets, where it is not possible to perform segmentation due to the different setup used for capture. The artefacts produced by the fusion is partially mitigated in the 3DHSv results, however, due to the resolution used, some finer details are lost. Finally, the results from 3DHSm present an accurate reproduction of minute object details, such as the ears of the *Corgi*, the ornaments on the neck of the *Llama* or the patterns on the neck of the *Giraffe*. In the *Fish* scene, the small cavity inside the eye is maintained as well as the features on the side of the tail fin and the two small concavities behind the dorsal fin. The high-frequency details of the scales on the side of the object are also faithfully reconstructed, while they are overly smoothed in the previous approaches. In the *Dragon* scene, the scales pattern on the side is accurately reconstructed and more details can be seen on the face compared to the other approaches.

We also perform a quantitative evaluation based on laser scans of *Fox*, *Corgi* and *Llama*, which are the only three fully rigid objects suitable for use as ground truth. The other three objects could not be used as they were either made of deformable materials (rubber in the case of *Duck* and *Giraffe*) or contained moving parts (moving wheels and axles in the case of *Bee*). Due to their specular nature, the objects had to be spray-coated with talc powder in order to enable laser scanning. The ground truth meshes were then manually registered to the set of images, followed by ICP alignment against the reconstruction for each approach in order to compute the geometric accuracy, normal accuracy and completeness metrics. Figure 21 shows the error maps for each method. The geometric accuracy, normal accuracy and completeness results are represented in Fig. 22 and Table 2. Results indicate some good performance overall, with the different methods achieving close to mm accuracy and a high completeness. Not surprisingly the VDo approach performs the worst out of the

four proposed HS methods. The VDp approach was found to be overall the best performer followed by 3DHSv, and then 3DHSm. It may be that the volumetric and mesh-based approaches are more sensitive to calibration errors than the view-dependent approach which would explain their lower performance on this dataset.

Finally, a run-time analysis is performed to benchmark the different methods. All experiments were performed on the same server with 16 cores and no use of GPU. Results are reported in Table 3. The view-dependent techniques are the slowest, requiring optimisation to be performed separately for each viewpoint. While this allows some parallelisation, this also generates some redundant processing, particularly if a large number of viewpoints are considered. The mesh-based approach is the fastest, while at the same time having a considerably lower memory footprint than the volumetric approach. None of the approaches have been optimised for speed and they all run entirely on the CPU. It is likely that performance could further reduced with some optimisation including porting of some of the operations to the GPU.

4.4 Comparison Against Multi-view Reconstruction Approaches

The proposed approach is compared against three 3D reconstruction techniques (one classical, two learning-based) on the real dataset released with this paper. The first approach, referred to as COLMAP, is based on Schönberger and Frahm (2016) and Schönberger et al. (2016). It is a classical approach consisting in jointly estimating depth and normal information by performing pixelwise view selection thanks to photometric and geometric priors and then minimising a multi-view geometric consistency term. This technique is one of the top performers on the Middlebury datasets. The second approach, referred to as CasMVSNet, is a deep learning approach, which is a fusion of MVSNet (Yao et al., 2018, 2019) and a technique that uses a cascade cost volume (Gu et al., 2020) to achieve higher resolution outputs during reconstruction. The third approach, referred to as NeuS, is a recent learning-based approach that combines a neural scene representation with a volume rendering technique (Wang et al., 2021).

To ensure a fair comparison, all the objects in the dataset were re-acquired under constant illumination conditions with the same number of views and similar camera placement as for the HS dataset. For the multi-view dataset, a constant ambient illumination was used instead of flash illumination and a camera was moved around the object instead of rotating the object using a turntable. This guarantees that the illumination conditions remain unchanged during the acquisition process, to ensure objects are acquired under optimal conditions for these methods. The results obtained on the real scenes are shown in Fig. 18 with error maps shown in Fig. 21. Figure 22 and Table 2 show the geometric accuracy, normal accuracy and completeness results. Note that for a fair comparison, prior to computing the error maps and error metrics for these methods, the reconstructions were truncated to remove any protrusions located below the ground plane level.

As it can be observed, COLMAP and CasMVSNet often produce poor reconstructions of the objects. The most notable examples for COLMAP are the *Fox* and *Llama* scenes, where untextured regions on the objects result in holes. This is

Table 2 Results obtained on the real datasets *Fox*, *Corgi* and *Llama*

	Method	G. Acc. ↓	N. Acc. ↓	Comp. ↑
Fox	VH	3.77	52.99	46.9
	3DHSc	4.24	73.79	50.0
	VDo	2.19	37.54	70.1
	VDp	**0.78**	**22.25**	**86.2**
	3DHSv	1.06	34.09	84.1
	3DHSm	1.38	33.50	72.1
	COLMAP	**1.71**	70.17	**69.1**
	CasMVSNet	N/A	N/A	N/A
	NeuS	2.15	**29.94**	57.6
Corgi	VH	3.92	46.12	41.0
	3DHSc	3.51	118.47	56.0
	VDo	1.74	49.03	82.9
	VDp	**0.79**	**17.54**	83.4
	3DHSv	1.10	29.15	**84.3**
	3DHSm	1.12	25.91	81.6
	COLMAP	**1.01**	40.99	**83.5**
	CasMVSNet	1.02	22.78	74.4
	NeuS	1.40	**17.96**	74.7
Llama	VH	3.76	44.08	30.0
	3DHSc	3.17	39.93	38.0
	VDo	2.18	89.51	**46.7**
	VDp	**1.94**	**25.62**	43.0
	3DHSv	2.06	29.72	40.0
	3DHSm	2.20	36.08	38.4
	COLMAP	1.94	78.93	37.7
	CasMVSNet	2.16	45.61	26.4
	NeuS	**1.90**	**22.25**	**44.9**

Geometric accuracy (expressed in mm and denoted by G. Acc.) and normal accuracy (expressed in degrees and denoted by N. Acc.) are computed at a 90% threshold, while the completeness (expressed in % and denoted by Comp.) is obtained at a threshold of 1 mm. Top performers are marked in bold, considering separately the HS approaches and the Multi-View Stereo (MVS) approaches

Table 3 Runtime comparison for the different methods benchmarked on the *Bee* dataset

	VDo	**VDp**	**3DHSv**	**3DHSm**
Time	4 h 29 m	52 h 45 m	12 h 46 m	1 h 30 m

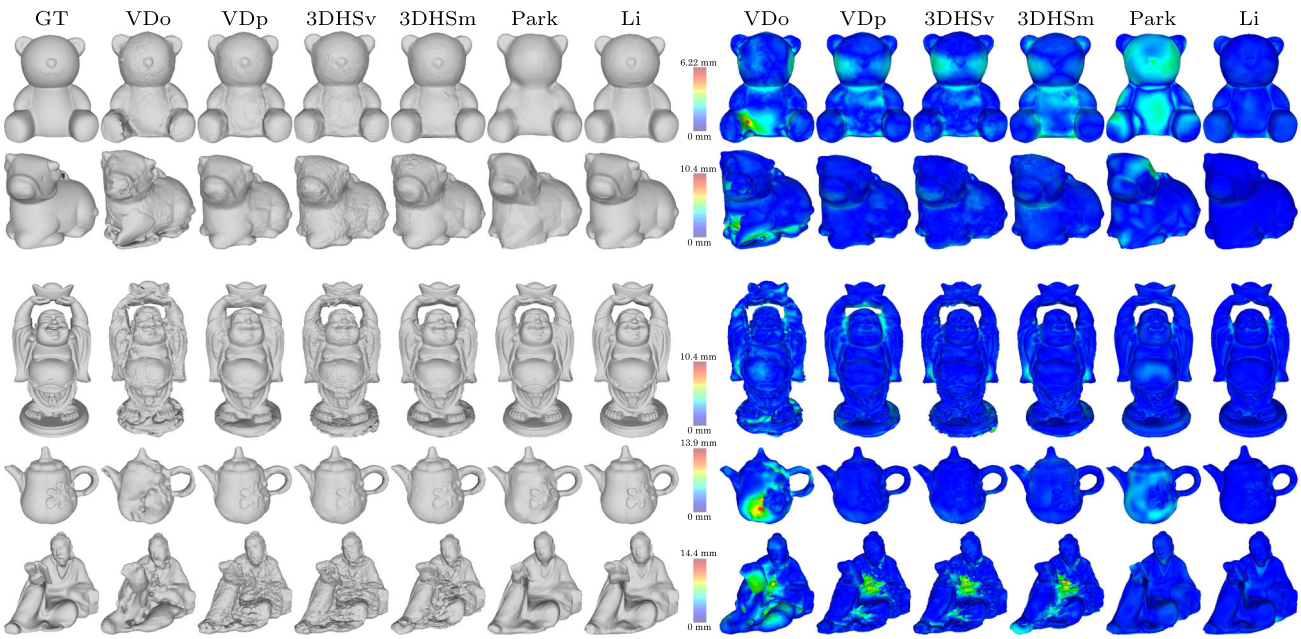

Fig. 23 Results on the DiLiGenT-MV datasets *Bear*, *Buddha*, *Cow*, *Pot2* and *Reading*

a result of point clouds being rather sparse and presenting holes that then lead to poor reconstruction once points are fused using Poisson surface reconstruction. The only exception is the *Giraffe* which, while being untextured, presents a Lambertian surface. Similarly, CasMVSNet produces poor reconstructions in the case of the *Fox*, *Duck*, *Llama* and *Giraffe*, obtaining very few point correspondences. This is likely due to the network not being trained on non-Lambertian scenes or due to the lack of textured surfaces. Although less dramatically affected, the other objects show very little detail and are noisy in places, ultimately resulting in worse results than the ones obtained using our proposed approaches. In contrast, NeuS was able to produce clean reconstructions, although these have a tendency to smooth out some of the detail. The quantitative analysis confirms that NeuS comes out as the top performer amongst the considered multi-view reconstruction approaches considered and comes close to the proposed methods.

4.5 Comparison Against Multi-View Photometric Stereo Approaches

The proposed approach is also evaluated on the DiLiGenT-MV dataset from Li et al. (2020). This dataset is tailored for reconstruction using MVPS approaches. It consists of five objects (*Bear*, *Buddha*, *Cow*, *Pot2* and *Reading*), each captured from 20 different viewpoints under 96 different illumination conditions, i.e. a total of 1920 images per object. To test our approaches, we used a subset of the dataset consisting of 20 pairs of images that were found to approximately satisfy the Helmholtz reciprocity constraint. More specifically, we defined these 20 pairs by considering all pairs of adjacent cameras, with View i lit by LED 20 paired with View $i+1$ lit by LED 68 in order to most closely approximate a reciprocal configuration. As such, our approaches are only able to use a fraction of the original dataset (approximately 2%).

We evaluate our two proposed view-dependent approaches VDo and VDp as well as the volumetric and mesh-based methods 3DHSv and 3DHSm. Performance is also compared against two MVPS approaches: the first one from Park et al. (2016) referred to thereafter as Park and the second one from Li et al. (2020) referred to as Li. Results for these two approaches are based on the mesh reconstructions provided for these two methods as part of the DiLiGenT-MV dataset. The reconstructed meshes for the different methods together with their corresponding error maps are shown in Fig. 23. Table 4 shows the geometric accuracy, normal accuracy and completeness metrics.

The proposed approaches based on Helmholtz reciprocity are able to achieve good reconstruction results, recovering generally well the geometric detail of the objects, even in the presence of complex topologies such as in the case of the *Buddha* and *Pot2*. Errors in reconstruction are most prominent in the case of *Reading* which contains some concavities causing inter-reflections that violate the reciprocity assumption. Amongst these four variants, VDo performs the worst. In contrast, VDp performs particularly well, outperforming the other three proposed approaches on all objects except *Bear* in which case 3DHSm is the top performer. This trend is confirmed by a qualitative analysis of the results which shows that VDp and 3DHSm produce more realistic results

Table 4 Results on the DiLiGenT-MV datasets *Bear*, *Buddha*, *Cow*, *Pot2* and *Reading*

	Method	G. Acc. ↓	N. Acc. ↓	Comp. ↑
Bear	VDo	2.30	29.54	65.1
	VDp	2.02	22.99	**67.0**
	3DHSv	2.06	22.82	65.4
	3DHSm	**1.48**	**19.34**	64.5
	Park	1.90	21.28	62.7
	Li	**0.39**	**6.93**	**75.3**
Cow	VDo	3.68	70.36	60.3
	VDp	**0.67**	**13.04**	**76.4**
	3DHSv	1.37	22.89	74.0
	3DHSm	1.21	20.44	77.2
	Park	0.89	25.56	**89.2**
	Li	**0.14**	**5.12**	77.2
Buddha	VDo	2.73	105.81	77.9
	VDp	**1.07**	**43.81**	74.3
	3DHSv	1.25	57.33	**85.9**
	3DHSm	1.15	44.63	82.4
	Park	1.95	36.16	**87.4**
	Li	**0.40**	**24.49**	86.2
Pot2	VDo	3.66	42.28	53.2
	VDp	**1.03**	**17.80**	**72.3**
	3DHSv	1.85	20.43	71.2
	3DHSm	1.43	20.21	70.0
	Park	2.70	20.72	37.7
	Li	**0.37**	**10.25**	**88.4**
Reading	VDo	5.63	96.14	57.8
	VDp	**2.40**	**52.00**	**80.4**
	3DHSv	2.78	62.97	78.1
	3DHSm	2.85	61.74	74.8
	Park	1.67	24.58	77.3
	Li	**0.47**	**15.59**	**84.3**

Geometric accuracy (expressed in mm and denoted by G. Acc.) and normal accuracy (expressed in degrees and denoted by N. Acc.) are computed at a 90% threshold, while the completeness (expressed in % and denoted by Comp.) is obtained at a threshold of 1 mm. Top performers are marked in bold, considering separately the HS approaches and the Multi-View Stereo (MVS) approaches

than VDo and 3DHSv which suffer from more noisy reconstructions.

When compared against the MVPS approaches, our proposed approaches are found to perform less well on this dataset. This is not surprising considering that they only utilise about 2% of the data the MVPS methods have access to. It is worth noting that our proposed approaches are still able to achieve reasonably close performance to Li, and are even outperforming the earlier method of Park et al. These results demonstrate the potential of the proposed approach as an alternative paradigm to model scenes from multi-view data under different illumination conditions. However, further evaluation is necessary to better understand how it compares against MVPS approaches under some more controlled evaluation conditions where each class of methods has access to the same amount of data.

5 Conclusions and Future Work

This paper introduced a family of BRDF-agnostic approaches for full 3D reconstruction. The first approach is based on fusing a set of orthographic or perspective view-dependent reconstructions from viewpoints distributed around the object. The second approach casts instead the problem as a volumetric optimisation problem seeking the optimal surface location within a voxel grid. The third approach uses a mesh-based formulation optimising vertices positions for a given mesh topology. Further, the paper contributes novel datasets to allow future benchmarking of full 3D HS approaches. An extensive experimental evaluation demonstrates that the proposed approaches are able to accurately model objects with complex materials, achieving sub-millimetre accuracy on the synthetic scenes and exhibiting robustness to image noise. A comparison against multi-view reconstruction techniques shows how the proposed HS approaches are able to improve reconstruction quality on challenging non-Lambertian low-texture objects where MVS approaches typically perform poorly. The proposed approaches were also compared against MVPS approaches and found to achieve good quality results while using only a fraction of the number of images.

The proposed approaches all suffer from the following limitations. Firstly, they are unable to handle scenes with significant subsurface scattering as these cannot be described by a BRDF which requires light to enter and leave a surface at the same point. In practice, we have observed that the approach is still able to reconstruct scenes exhibiting small amounts of subsurface scattering such as in the case of the *Duck* scene in our dataset. Secondly, performance degrades in the presence of inter-reflections, particularly in concavities where these are likely to be significant such as in the case of the *Reading* scene from the DiLiGenT-MV dataset. Finally, the approach usually fails in the presence of highly reflective surfaces (mirror-like surfaces) due to the limited dynamic range of the sensors used during acquisition; this may be alleviated by making use of high dynamic range imaging.

An interesting avenue for future work would be to extend the approach to enable capture in less controlled environments such as outdoor settings. This could be achieved in principle by capturing an additional image acquired with only the ambient illumination which could then be factored out from the images acquired using the light sources for Helmholtz reciprocal pair acquisition. Another avenue for future work would be to extend the approach to dynamic

scenes. To date, use of HS for dynamic scenes has been limited to 2.5D reconstruction. The extension to full 3D could be made possible via multi-spectral imaging using a larger number of frequency bands or using temporal multiplexing. Finally, another direction would be to explore the use of deep neural networks. This may prove beneficial in overcoming the remaining limitations relating to dealing with subsurface scattering and inter-reflections which are currently difficult to model explicitly.

Acknowledgements The authors would like to thank *3D Enterprise* for producing the laser scans for the *Fox*, *Corgi* and *Llama* objects.

Funding This work was supported by the UKRI EPSRC Doctoral Training Partnership Grant EP/N509772/1 (studentship reference 1815219), the UKRI EPSRC Research Grant EP/M021793/1, the UKRI EPSRC Audio-Visual Media Research Platform Grant EP/P022529/1 and the Royal Society Research Grant RG150625.

Availability of data The authors confirm that the datasets generated as part of this research are freely available under the terms and conditions detailed in the licence 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.900004 and the data repository website: https://cvssp.org/data/HS_DS/.

Appendix A: Confidence Scores Ablation

This section conducts further analysis to evaluate the effect of the confidence scores that were introduced in the context of the view-dependent approaches in Sect. 3.1. To this end, an ablation study is conducted for both the VDo and the VDp approaches. In each case, we compare the reconstructions obtained with no confidence score (denoted by *no conf.*), only the saliency-driven confidence score (denoted by C_s), only the surface normal-driven confidence score (denoted by C_n) or the complete confidence score combining saliency-driven and surface normal driven scores (denoted by C_{sn}). Quantitative results obtained on the synthetic dataset using geometric accuracy, normal accuracy and completeness are reported in Table 5. These indicate the benefit of using the combined confidence scores which consistently outperforms

Table 5 Results on synthetic scenes using the view-dependent methods with different confidence scores: no confidence (denoted by *no conf.*), saliency-driven confidence only (denoted by C_s), surface normal-driven confidence only (denoted by C_n) and complete confidence score combining salience-driven and surface normal-driven scores (denoted by C_{sn})

	Method	No Noise			0.01% Noise			0.1% Noise		
		G. Acc. ↓	N. Acc. ↓	Comp. ↑	G. Acc. ↓	N. Acc. ↓	Comp. ↑	G. Acc. ↓	N. Acc. ↓	Comp. ↑
Armadillo	VDo – *no conf.*	0.26	18.62	97	0.26	18.50	97	0.26	18.58	97
	VDo – C_s	0.26	18.62	97	0.26	18.51	97	0.26	18.57	97
	VDo – C_n	0.26	18.61	97	0.26	18.51	97	0.26	18.57	97
	VDo – C_{sn}	0.26	18.61	97	0.26	18.51	97	0.26	18.57	97
	VDp – no conf	0.19	15.43	98	0.20	15.93	97	0.21	15.94	97
	VDp – C_s	0.18	15.24	99	0.19	15.71	**98**	0.20	15.80	**98**
	VDp – C_n	**0.17**	15.09	98	**0.18**	15.52	**98**	**0.19**	15.51	**98**
	VDp – C_{sn}	**0.17**	**15.01**	**99**	**0.18**	**15.44**	**98**	**0.19**	**15.49**	**98**
Bunny	VDo – *no conf.*	0.49	13.22	91	0.52	12.60	90	0.57	12.55	88
	VDo – C_s	0.49	13.22	91	0.52	12.59	90	0.57	12.56	88
	VDo – C_n	0.49	13.22	91	0.52	12.59	90	0.57	12.56	88
	VDo – C_{sn}	0.49	13.22	91	0.52	12.59	90	0.57	12.56	88
	VDp – no conf	0.24	7.26	97	0.28	7.57	96	0.29	7.74	95
	VDp – C_s	0.23	7.08	97	0.26	7.36	96	0.28	7.62	**96**
	VDp – C_n	0.23	7.18	97	0.27	7.50	96	0.28	7.73	**96**
	VDp – C_{sn}	**0.22**	**7.02**	**98**	**0.25**	**7.28**	**97**	**0.27**	**7.56**	**96**

The analysis is conducted under different noise levels. Geometric accuracy (expressed in mm and denoted by G. Acc.) and normal accuracy (expressed in degrees and denoted by N. Acc.) are computed at a 90% threshold, while the completeness (expressed in % and denoted by Comp.) is obtained at a threshold of 0.5 mm. Top performers are marked in bold

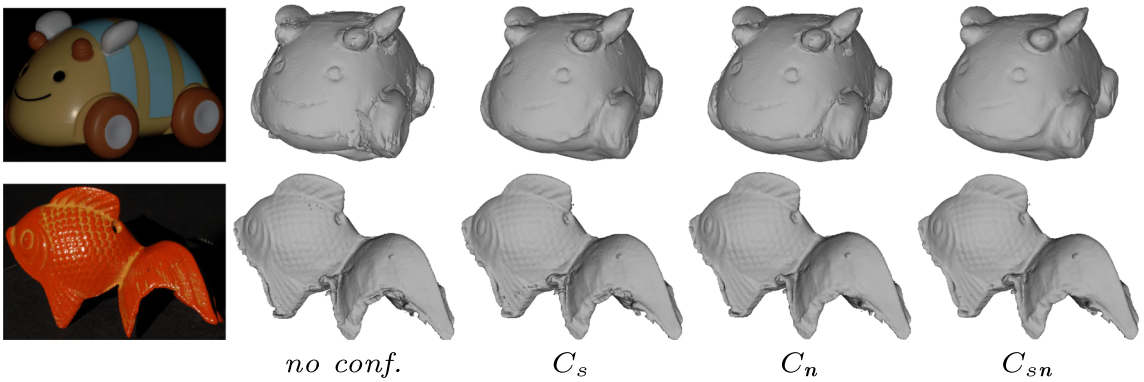

Fig. 24 Results obtained on the real dataset *Bee* and *Fish* using the perspective view-dependent HS approach VDp with different confidence scores considered during the fusion process

all other approaches across all three metrics in the case of the VDp approach. The use of confidence score does not appear to yield any improvement in the case of the VDo approach. This is due to the fact there is limited overlap between view-dependent reconstructions with this approach which only considers six views. In contrast, the benefit is more apparent with the VDp approach which uses a larger number of overlapping views. Qualitative results obtained using the VDp approach are illustrated in the case of the *Bee* and *Fish* objects in Fig. 24. These show a modest yet noticeable improvement, confirming the benefit of incorporating confidence into the fusion process.

References

Ackermann, J., & Goesele, M. (2015). A survey of photometric stereo techniques. *Foundations and Trends in Computer Graphics and Vision, 9*(3–4), 149–254.

Addari, G., & Guillemaut, J.-Y. (2019a). An MRF optimisation framework for full 3D Helmholtz stereopsis. In *International conference on computer vision theory and applications* (pp. 725–736).

Addari, G., & Guillemaut, J.-Y. (2019b). Towards globally optimal full 3D reconstruction of scenes with complex reflectance using Helmholtz stereopsis. *European conference on visual media production*.

Addari, G., & Guillemaut, J.-Y. (2020). An MRF optimisation framework for full 3D reconstruction of scenes with complex reflectance. In *Computer vision, imaging and computer graphics theory and applications. VISIGRAPP 2019. CCIS* (Vol. 1182, pp. 456–476). Berlin: Springer.

Audenaert, E. A., Houcke, J. V., Almeida, D. F., Paelinck, L., Peiffer, M., Steenackers, G., & Vandermeulen, D. (2019). Cascaded statistical shape model based segmentation of the full lower limb in CT. *Computer Methods in Biomechanics and Biomedical Engineering, 22*(6), 644–657.

Baumgart, B. G. (1974). Geometric modeling for computer vision (Technical Report). Computer Science Department, Stanford University.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological), 48*(3), 259–302.

Bi, S., Xu, Z., Sunkavalli, K., Hašsan, M., HoldGeoffroy, Y., Kriegman, D., & Ramamoorthi, R. (2020). Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *European Conference on Computer Vision* (pp. 294–311).

Boyer, E., & Franco, J.-S. (2003). A hybrid approach for computing visual hulls of complex objects. In *Computer vision and pattern recognition*.

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 *European Conference on Computer Vision* (pp. 628–644).

Cipolla, R., & Blake, A. (1992). Surface shape from the deformation of apparent contours. *International Journal of Computer Vision, 9*, 83–112.

Dai, Y., Zhu, Z., Rao, Z., & Li, B. (2019). MVS2: Deep unsupervised multi-view stereo with multi-view symmetry. In *International conference on 3d vision*.

Delaunoy, A., Prados, E., & Belhumeur, P. N. (2010). Towards full 3D Helmholtz stereovision algorithms. In *Asian conference on computer vision* (pp. 39–52).

Forbes, K., Voigt, A., & Bodika, N. (2004). Visual hulls from single uncalibrated snapshots using two planar mirrors. In *South African workshop on pattern recognition*.

Fyffe, G., Graham, P., Tunwattanapong, B., Ghosh, A., & Debevec, P. (2016). Near-instant capture of high-resolution facial geometry and reflectance (Vol. 35) (No. 2).

Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., & Debevec, P. (2011). Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics*. https://doi.org/10.1145/2070752.2024163

Gilbert, A., Volino, M., Collomosse, J., & Hilton, A. (2018). Volumetric performance capture from minimal camera viewpoints. In *European conference on computer vision* (pp. 591–607).

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Computer vision and pattern recognition* (pp. 2492–2501).

Guillemaut, J.-Y., Drbohlav, O., Illingworth, J., & Sára, R. (2008). A maximum likelihood surface normal estimation algorithm for Helmholtz stereopsis. In *International conference on computer vision theory and applications* (pp. 352-359).

Guillemaut, J.-Y., Drbohlav, O., Sára, R., & Illingworth, J. (2004). Helmholtz stereopsis on rough and strongly textured surfaces. In *International symposium on 3D data processing, visualization and transmission* (pp. 10–17).

Han, T.-Q., & Shen, H.-L. (2015). Photometric stereo for general BRDFs via reflection sparsity modeling. *IEEE Transactions on Image Processing, 24*(12), 4888–4903.

Holroyd, M., Lawrence, J., & Zickler, T. (2010). A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance. *ACM Transactions on Graphics, 29*(4), 1–12. https://doi.org/10.1145/1778765.1778836

Ikehata, S. (2018). CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In *European conference on computer vision* (pp. 3–19).

Janko, Z., Drbohlav, O., & Sára, R. (2004). Radiometric calibration of a Helmholtz stereo rig. In *Computer vision and pattern recognition*.

Kar, A., Tulsiani, S., Carreira, J., & Malik, J. (2015). Category-specific object reconstruction from a single image. In *Computer vision and pattern recognition* (pp. 1966–1974).

Kazhdan, M., Bolitho, M., & Hoppe, H. (2006). Poisson surface reconstruction. In *Eurographics symposium on geometry processing* (pp. 61-70).

Kolmogorov, V. (2006). Convergent treereweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1568–1583.

Kolmogorov, V. (2015). A new look at reweighted message passing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(5), 919–930.

Kolmogorov, V., & Rother, C. (2007). Minimizing nonsubmodular functions with graph cuts: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(7), 1274–1279.

Krishnamurthy, V., & Levoy, M. (1996). Fitting smooth surfaces to dense polygon meshes. In *Annual conference on computer graphics and interactive techniques* (pp. 313–324).

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(2), 150–162.

Lewis, R. R. (1994). Making shaders more physically plausible. *Computer Graphics Forum, 13*(2), 109–120.

Li, M., Zhou, Z., Wu, Z., Shi, B., Diao, C., & Tan, P. (2020). Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing, 29*, 4159–4173.

Liang, C., & Wong, K.-Y.K. (2010). 3D reconstruction using silhouettes from unordered viewpoints. *Image and Vision Computing, 28*(4), 579–589.

Liu, X., Yao, H., Yao, G., & Gao, W. (2006). A novel volumetric shape from silhouette algorithm based on a centripetal pentahedron model. In *International conference on pattern recognition*.

Logothetis, F., Budvytis, I., Mecca, R., & Cipolla, R. (2020). A CNN based approach for the nearfield photometric stereo problem. In *British machine vision conference*.

Logothetis, F., Mecca, R., & Cipolla, R. (2019). A differential volumetric approach to multiview photometric stereo. In *International conference on computer vision* (pp. 1052–1061).

Lombardi, S., & Nishino, K. (2016). Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. In *International conference on 3D vision* (pp. 305–313).

Ma, W.-C., Hawkins, T., Peers, P., Chabert, C.-F., Weiss, M., & Debevec, P. (2007). Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Eurographics conference on rendering techniques* (pp. 183–194).

Magda, S., Kriegman, D. J., Zickler, T. E., & Belhumeur, P. N. (2001). Beyond Lambert: Reconstructing surfaces with arbitrary BRDFs. In *International conference on computer vision* (pp. 391–398).

Mallick, S. P., Zickler, T. E., Kriegman, D. J., & Belhumeur, P. N. (2005). Beyond Lambert: Reconstructing specular surfaces using color. In *Computer vision and pattern recognition* (pp. 619–626).

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J., & McMillan, L. (2000). Image-based visual hulls. In *Annual conference on computer graphics and interactive techniques* (pp. 369–374).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99–106.

Nicodemus, F.E., Richmond, J.C., Hsia, J.J., Ginsberg, I.W., Limperis, T. (1977). *Geometrical considerations and nomenclature for reflectance*. Monograph 160, National Bureau of Standards. Washington D.C. https://doi.org/10.6028/NBS.1412MONO.160

Oechsle, M., Peng, S., & Geiger, A. (2021). UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International conference on computer vision* (pp. 5589–5599).

Oxholm, G., & Nishino, K. (2014). Multiview shape and reflectance from natural illumination. In *Computer vision and pattern recognition* (pp. 2163–2170).

Park, J., Sinha, S. N., Matsushita, Y., Tai, Y.-W., & Kweon, I. S. (2016). Robust multiview photometric stereo using planar mesh parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(8), 1591–1604.

Roubtsova, N., & Guillemaut, J.-Y. (2014a). A bayesian framework for enhanced geometric reconstruction of complex objects by Helmholtz stereopsis. In *International conference on computer vision theory and applications* (Vol. 3, pp. 335–342).

Roubtsova, N., & Guillemaut, J.-Y. (2014b). Colour Helmholtz stereopsis for reconstruction of complex dynamic scenes. In *International conference on 3D vision* (pp. 251–258).

Roubtsova, N., & Guillemaut, J.-Y. (2017). Colour Helmholtz stereopsis for reconstruction of dynamic scenes with arbitrary unknown reflectance. *International Journal of Computer Vision, 124*(1), 18–48.

Roubtsova, N., & Guillemaut, J.-Y. (2018). Bayesian Helmholtz stereopsis with integrability prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(9), 2265–2272.

Schönberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. In *Computer vision and pattern recognition* (pp. 4104–4113).

Schönberger, J. L., Zheng, E., Pollefeys, M., & Frahm, J.-M. (2016). Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision* (pp. 501–518).

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. *Computer vision and pattern recognition* (pp. 519–528).

Snyder, W. C. (2002). Structured surface BRDF reciprocity: Theory and counterexamples. *Applied Optics, 41*(21), 4307–4313.

Szeliski, R. (1993). Rapid octree construction from image sequences. *CVGIP: Image Understanding, 58*, 23–32.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., & Rother, C. (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.

Tarini, M., Callieri, M., Montani, C., Rocchini, C., Olsson, K., & Persson, T. (2002). Marching intersections: An efficient approach to shape-from-silhouette. In *Conference on vision, modeling, and visualization* (pp. 283–290).

Tu, P., & Mendonca, P. R. S. (2003). Surface reconstruction via Helmholtz reciprocity with a single image pair. *Computer vision and pattern recognition*.

Tunwattanapong, B., Fyffe, G., Graham, P., Busch, J., Yu, X., Ghosh, A., & Debevec, P. (2013). Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Transactions on Graphics, 32*(4), 1–12.

Turk, G., & Levoy, M. (1994). Zippered polygon meshes from range images. In *Annual Conference on Computer Graphics and Interactive Techniques* (pp. 311–318).

Vogiatzis, G., Hernández Esteban, 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.

von Helmholtz, H. (1924). *Helmholtz's treatise on physiological optics*, Vol. I (J.P.C. Southall, Trans.). Optical Society of America. https://doi.org/10.1037/13536-000

Wainwright, M. J., Jaakkola, T. S., & Willsky, A. S. (2005). MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions on Information Theory, 51*(11), 3697–3717.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., & Wang, W. (2021). NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems, 34*, 27171–27183.

Ward, G. J. (1992). Measuring and modeling anisotropic reflection. *SIGGRAPH Computer Graphics, 26*(2), 265–272.

Weinmann, M., Ruiters, R., Osep, A., Schwartz, C., & Klein, R. (2012). Fusing structured light consistency and Helmholtz normals for 3D reconstruction. In *British machine vision conference*.

Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical Engineering*. https://doi.org/10.1117/12.7972479

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). MVSNet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision* (pp. 785–801).

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent MVSNet for highresolution multi-view stereo depth inference. In *Computer vision and pattern recognition* (pp. 5525–5534).

Yariv, L., Gu, J., Kasten, Y., & Lipman, Y. (2021). Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems, 34*, 4805–4815.

Zhang, K., Luan, F., Wang, Q., Bala, K., & Snavely, N. (2021). PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Computer vision and pattern recognition* (pp. 5453–5462).

Zickler, T. E. (2006). Reciprocal image features for uncalibrated Helmholtz stereopsis. In *Computer vision and pattern recognition* (pp. 1801–1808).

Zickler, T. E., Belhumeur, P. N., & Kriegman, D. J. (2002). Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. *International Journal of Computer Vision, 49*(2), 215–227.

Zickler, T. E., Ho, J., Kriegman, D. J., Ponce, J., & Belhumeur, P. N. (2003). Binocular Helmholtz stereopsis. In: *International Conference on computer vision* (pp. 1411–1417).