3D reconstruction of insects: an improved multifocus stacking and an evaluation of learning-based MVS approaches

Chang Xu, Jiayuan Liu, Chuong Nguyen, Fabien Casten, Benoit Maujean, Simone Gasparini

3D reconstruction of insects: an improved multifocus stacking and an evaluation of learning-based MVS approaches

Chang Xu*, Jiayuan Liu*
Australian National University
Canberra, Australia
{u5756756, u6409573}@anu.edu.au

Chuong Nguyen
CSIRO Data61
Canberra, Australia
chuong.nguyen@csiro.au

Fabien Casten, Benoit Maujean
Technicolor
Paris, France
{fabien.castan, benoit.maujean}@technicolor.com

Simone Gasparini
University of Toulouse – IRIT
Toulouse, France
simone.gasparini@irit.fr

Abstract

3D reconstruction of insects from photographs is a challenging task as it requires to tackle several problems such as strong out-of-focus areas in macro-photography, thin structures (insect legs and hairs), flat-colored surfaces (insects shells), non-Lambertian (shells specularities) and even translucent surfaces (wings). In this work, we first present a new lens-based image registration technique for accurate multi-focus stacking suitable for 3D reconstruction purposes while other methods create in-focus images for viewing purpose only. We then evaluate and compare the classical Multi-View-Stereo (MVS) reconstruction pipeline for small and complex objects with recent deep learning-based reconstruction methods such as the Neural Radiance Fields (NeRF) and the Neural Sparse Voxel Fields (NSVF). We present an assessment of different sources of errors for the considered methods. The results are compared both quantitatively and qualitatively across the different methods. From this analysis we present a series of practical guidelines for addressing the common issues of the reconstruction of small objects under challenging conditions.

1. Introduction

Traditional multiview stereo (MVS) reconstruction [21] has become a popular tool for image-based 3D reconstruction and readily available as in a number of open-source and commercial software [20, 11, 2, 3, 19, 1]. MVS by itself has shown to be sufficiently accurate for large scale scenes and objects [10, 12]. However, 3D reconstruction of insects and biological specimens is challenging due to small sizes, fine features, complex surface properties, and transparency. These problems have been tackled by multiple works [17, 22, 18] by extending MVS to work with multifocus images. Multifocus multiview stereo (MMVS) has been developed and applied with significant success to create true-color 3D models of challenging specimens:

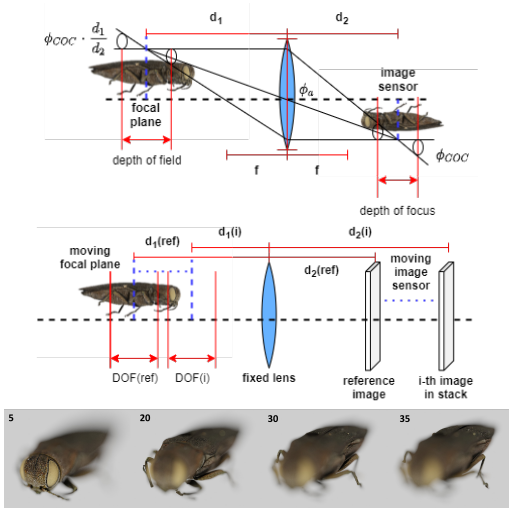


Figure 1: Top: lens image formation and circle of confusion. Middle: multi-focus capture with fixed lens setup. Bottom: example images from one stack of 61 images.

- Small size (a few mm to a few cm long) is overcome by using a high magnification lens. As such lens has narrow depth of focus, a set of multiple partially-focus images are captured with a focal plane moving across the whole depth of the object as shown in Fig. 1. These images are then registered and "stacked" to produce a

*These authors contributed equally to the work.

synthetically in-focus image for a single view. Repeating the same process for multiple views [17] to be used for 3D reconstruction. Careful camera calibration for multifocus image registration improved the accuracy of 3D reconstruction [22]. The issue of perspective inconsistency was discovered and solved by a fixed lens setup [13, 6].

- Specular reflection interferes with stereo matching and cause significant error. It could be reduced by dome illumination [22] or polarisation filter.
- Fine and thin structure is another challenge for pose and depth estimation of MVS. Backlight illumination and automatic masking and Visibility-Consistent Meshing (Agisoft Photoscan [3]) were used to improve depth estimation to some extent [22]. Hair-like structures and surfaces are still a challenge.
- Transparency is the most challenging problem, often leading to incomplete 3D reconstructed models. Such 3D models need to be fixed manually with post-processing using 3D editing software [18].

These issues remain partially and continue to be an obstacle to more realistic 3D modeling of objects. "Out-of-the-box" solutions could open to new possibilities including a) ground-truth or prior information for camera poses and image registration, b) deep-learning based 3D reconstruction. Using accurate motorised control could provide known poses for captured images, therefore potentially removing a major source of error. Recent progress in deep-learning based 3D modeling including NeRF [15] and NSVF [14] also leads to better modeling complex geometries and surfaces.

The paper proposes three main contributions. We present (i) a new lens-based image registration technique (*c.f*. Section 2) for accurate multifocus stacking (*c.f*. Section 3.2) suitable for 3D reconstruction while other methods create in-focus images for viewing purpose only. We then present (ii) an extensive assessment of different sources of errors (*c.f*. Section 3) for the 3 selected methods, MVS, NeRF and NSVF. The results are compared and validated both quantitatively and qualitatively across the different methods. From this analysis, we present (iii) a series of guidelines (*c.f*. Section 4) for tackling the major challenges that arise when reconstructing small object under challenging conditions. Code and data used in this paper are released in open source[1].

2. Methodology

2.1. Multi-focus capture and stacking

Image capturing with fixed lens. For conventional multi-focus image capturing, the lens and camera moves together, so as the lens move, the distance and scale of different parts of the object changes differently, leading to perspective distortion. To overcome this issue, fixed-lens [13, 6] is used where the camera lens is fixed, only the image sensor moves step by step when capturing images as shown by Fig. 1. The image formation follows the lens equation $d_2 = \frac{d_1 f}{d_1 - f}$, where d_2 represents distance from camera lens to image sensor, d_1 represents distance from camera lens to focal plane, and f represents focal length.

The camera needs to change the position of its focal plane such that the overlap of multiple depth of fields (DOF), *i.e*. the distance range at which parts of the object are in focus, covers the full depth of the specimen. It is determined by the circle of confusion with diameter ϕ_{COC} which describes the point spread of a light source directed onto the focal plane by the lens. An image is considered in focus as long as the size of the point formed on the image plane is smaller than ϕ_{COC} (usually chosen to be 0.1% of the image width). The relation between DOF and ϕ_{COC} is illustrated by Fig. 1 and given by $DOF = 2\frac{d_1 f \phi_{COC}}{(d_1 - f)\phi_a}$, where ϕ_a is the diameter of lens aperture. By moving image sensor, d_2 changes and so does d_1 without changing perspective projection. For a specimen with max depth length D, the number of images to capture is $N = \frac{D}{DOF}$ with the moving step $\Delta d_2 = d_2 - \frac{(d_1 + DOF)f}{(d_1 + DOF) - f}$.

Image alignment. Before focus stacking, the images are aligned by homography transformation between one image to a reference image. Homography transformation can be computed using directly matching the images of the specimen. To improve the estimation accuracy of homography transformation, a calibration target could be used. A dual calibration target forming an angle of 90° (*c.f*. Supplementary material) was used to allow larger depth range.

Focus stacking. To stack the multi-focus images, they were firstly aligned to a reference image in the stack by using the computed homography matrices. Once the images were aligned, their in-focus regions were fused by computing the weight using Laplacian of Gaussian Pyramid [24].

Background removal. The image capturing process was done in frontlight and backlight environment. The backlight images served as the masks to remove the background noise in frontlight images. The background removal was done by firstly converting the backlight images to binary images that outlines the object's silhouette, and performing bit-wise operations with the frontlight images.

2.2. Image alignment method using lens equation

For fixed-lens image capturing, the image sensor only translates back and forth, so captured images only experience scaling changes:

$$S(i) = \frac{d_2(ref)}{d_2(i)} \tag{1}$$

[1] https://github.com/chuong/3d_insect_recon_validation

where $d_2(ref)$ and $d_2(i)$ are the distances from the lens to the image sensor for a reference image and an arbitrary image i.

To account for the optical centre in the middle of the image, an offset from the top-left corner to the middle is also included. The homography for fixed-lens setup becomes:

$$H(i) = \begin{bmatrix} S(i) & 0 & \frac{1}{2}w(-S(i)+1) \\ 0 & S(i) & \frac{1}{2}h(-S(i)+1) \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

where w and h are the image width and height.

The affine homography transformation given by Eq. (2) for focus stacking for fixed lens camera setup is a special case of the one described in [6]. Equation (2) is a simpler and more practical solution with the assumption that camera sensor is aligned well enough with the macro rail.

2.3. Image-based reconstruction methods

Multiview stereo reconstruction (MVS). MVS was performed using the open-source 3D reconstruction software Meshroom [11]. It provides a graphic interface that enables the user to customise a reconstruction pipeline, execute the reconstruction, and visualize estimated camera poses, dense point cloud and resulting 3D models. The 3D reconstruction pipeline for Meshroom can be summarized as follows: structure from motion to estimate camera poses and sparse point cloud, depth map estimation to generate dense point cloud, and meshing the dense point cloud and texturing the mesh.

Neural Radiance Fields (NeRF). NeRF is a multilayer perceptron (MLP) deep network [15] that trains on a set of multiview input images to synthesise new realistic view or extract 3D shape model. For each view, NeRF takes as input a 5D vector including the spatial location coordinate x, y, z, and viewing azimuthal angle θ and polar angle ϕ relative to the reconstructed volume. NeRF requires images, camera poses and camera intrinsics as inputs. Camera poses can be estimated using traditional structure from motion.

Neural Sparse Voxel Fields (NSVF). NSVF is an MLP-based reconstruction approach similar to NeRF, but it incorporates sparse voxel octree for efficient rendering training and rendering [14]. While NeRF allocates a fixed computational budget for every ray, NSVF only calculates the ray through only the sparse voxels. Therefore, NSVF can compute faster and more efficient than NeRF, theoretically over ten times faster. NSVF takes images, camera poses, camera intrinsics, and bounding box as inputs.

2.4. Validation strategies

In this paper, we examine the impact of all important sources of errors to the final 3D reconstruction quality of challenging objects. We also validate MVS, NeRF and NVSF approaches against each other.

Effect of different image registrations for multi focus stacking. This is important when dealing with small objects and multifocus stacking before performing 3D reconstruction. The alignment could rely on the information contained in the object images, calibration images, or image formation via lens equation.

Effect of distribution of camera poses when taking multiview images. As a pan-tilt motorised stage is often used to rotate the object or move the camera on a spherical surface, the distribution of the camera poses around the object is often non-uniform. Three strategies of pose distribution are examined in this paper.

Effect of image resolutions. Low-cost lenses often yield low optical resolution and this could put a constraint the effective image resolution and therefore the quality of the final 3D reconstructed models. This paper aims at investigating how critical image resolution is to different 3D reconstruction approaches.

Effect of prior pose estimation. All 3D reconstruction approaches require camera poses of the input images to perform. The camera poses could be estimated by structure from motion, or prior-calibrated pose by a pan-tilt motorised stage. This paper aims to find out how much ground truth poses improve the 3D reconstruction as compared to using estimated poses.

Effect of different object geometries and materials, including transparency and reflection. These are known to be challenging for MVS, studies of NeRF and NSVF suggest that they could handle these challenging problems.

To validate the accuracy of different alignment methods for focus stacking, Structure Similarity Index (SSIM) [25] was used. This metrics quantifies the difference between an image and a reference ground-truth image. The closer the score is to 1, the more similar the two images are.

Validations of 3D reconstruction accuracy are also performed by computing the Hausdorff distance (using MeshLab [7]) of 3D reconstructed meshes relative to the ground-truth 3D meshes. Hausdorff distance is defined as the largest distance between two meshes. So when comparing two meshes, large Hausdorff distances indicate missing features in the reconstruction results. Once Hausdorff distance was computed and normalised by the diagonal distance of the 3D mesh bounding box, it is visualised as heatmap on the ground truth meshes.

3. Results

3.1. Data generation and preprocessing

3D rendering and ground truth poses. The experiment used synthetic images of realistic 3D scanned models from [22, 8] rendered in Blender so that the 3D models could serve as the ground truth meshes to validate the effect of using different parameters in the 3D reconstruction

pipeline.

The image capturing environment was modelled using an open-source Blender add-on for photogrammetry partially based on [4]. The Blender add-on generates a vertical array of cameras using the settings user provided. Each camera rotates around the object and render a stack of multi-focus images using fixed-lens setup at multiple views. Unless mentioned elsewhere, the azimuth angle is modified such that the camera poses are approximately uniformly distributed around a sphere, with the centre being the object. The add-on was also modified to save the camera poses and convert to Meshroom format.

The image capturing process was applied to object in frontlight and backlight environment, where the backlight images could serve as masks to remove background noise from the frontlight images as in [22].

Image stacking and background removal. Focus stacking involve aligning the multi-focus images using their homography matrices, then fusing their in-focused regions. In the experiment, the multi-focus images of calibration target were stacked first to provide the calibrated homography matrices. The homography matrices were computed using different methods such as Enhanced Correlation Coefficient (ECC [9]) homography, ECC affine, *etc*., to study the effect of homography transformation on the reconstruction result. Once the images for the calibration target were stacked, the front-light and back-light images for the object were stacked using the same calibrated homography matrices.

The stacked back-light images were then used as masks to remove the background noises from the stacked front-light images. The background removal was achieved by firstly converting the stacked back-light images to binary images using a threshold, then bit-wise operation between the binary images and the stacked front-light images was performed to remove the background.

Pose estimation. In this paper, Meshroom [11] is chosen to perform structure from motion for pose estimation and multiview stereo reconstruction (MVS). As a result, from now on the term MVS will be used to refer to the reconstruction results produced by Meshroom.

3.2. Effect of homography transformations for image stacking

We tested 2 alignment methods for focus stacking, ECC affine homography (OpenCV [5]'s *findTransformECC* function based on [9]) and the theoretical transformation from lens equation. For methods using ECC affine homography, we also tested them under different conditions such as with and without calibration, and with and without linear fitting.

ECC affine homography was able to compute the transformation matrices. However, they produce distortions in stacking results, as shown by Fig. 3.

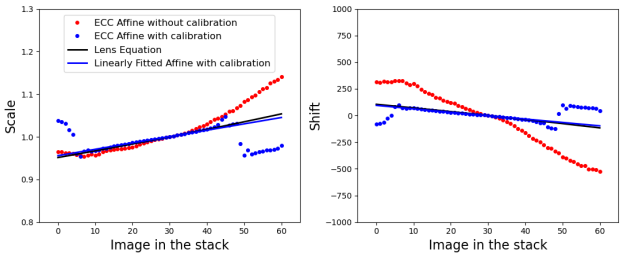


Figure 2: Distributions of the scale and shift in homography matrices estimated using different alignment methods.

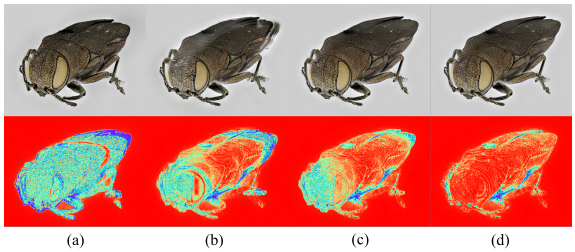


Figure 3: Focus stacked images and error maps. (a) ECC affine w/o calibration target. (b) & (c) ECC affine with calibration target and its linear fitting. (d) Lens equation.

The distribution of the scaling and shifting factor in ECC affine homography with and without calibration target shows that this method is unreliable near the two extrema of the image stack due to the large areas out of focus. To improve the result of standard ECC affine homography methods, linear fitting is applied to the middle portion of the data point. To quantify the accuracy of all alignment methods, the SSIM algorithm was applied to the output images and the ground truth image rendered without the out of focus effect. The results are shown in Fig. 3 and Table 1. It could be noted that while linear fitting was able to eliminate the distortions in standard calibrated ECC affine homography methods, it was not applicable to non-calibrated ECC affine homography methods since its scaling and shifting components were highly different to the theoretical trend. This difference is found to vary from one view to another.

Focus Stacking Methods	**SSIM Score**
ECC Affine w/o calibration target	0.938
ECC Affine with calibration target	0.971
Linearly Fitted ECC Affine w.c.t.	0.976
Our lens equation (Eq. (2))	0.983

Table 1: SSIM score for focus stacking methods.

Without using a calibration target for focus stacking, subsequent MVS reconstruction failed to produce reasonable results and therefore this case is ignored from now.

To validate the significance of different image alignments, reconstruction was performed using stacked images from ECC affine, linearly fitted ECC affine, and theoretical calculation. The results and their comparison to the ground truth model are shown by Fig. 4 and Table 2. As the results shown, the reconstruction result using stacked images from linearly fitted ECC affine was not significantly different to the reconstruction result using stacked images from theoretical calculation. On the other hand, the reconstruction result using stacked images from standard ECC affine was much worse, which suggests that distortions in focus stacking contribute significantly to the reconstruction result. Therefore, in practice, if theoretical calculation could not be achieved, it is recommended to linearly fit the standard ECC affine homography matrices.

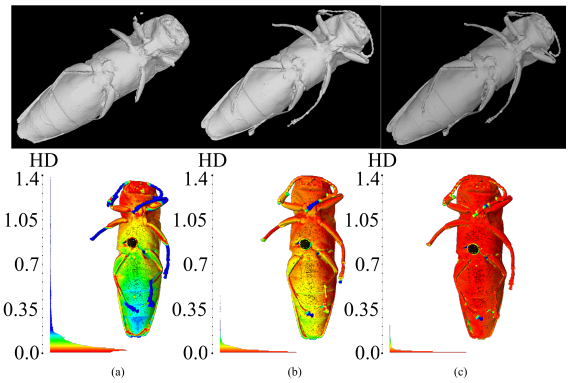


Figure 4: Reconstruction result using stacked images (top row) from different alignment methods and their Hausdorff distance to ground truth model (bottom row). (a) ECC affine with calibration target. (b) Linearly fitted ECC affine with calibration target. (c) Lens equation. Using estimated poses.

Alignment methods	**Normalised HD**
ECC Affine with calibration target	0.00922
Linearly fitted ECC Affine w.c.t.	0.00241
Our lens equation (Eq. (2))	0.00154

Table 2: Hausdorff distance (HD) normalised by bounding box diagonal for reconstruction results using different focus stacking methods.

3.3. Effect of image pose distribution to MVS

The conventional way to capture images is to keep the camera pointing to the object, and apply constant step for pan (azimuth) and tilt rotation (polar angle) leading to denser camera pose distribution at the poles than at the equator region. To reduce the non-uniform distribution, two new simple strategies are proposed: one with linear variation of number of poses from 36 to 1 along z rotation axis, another with linear variation of the number of poses with tilt angle. Figure 5 shows the 3 different pose distributions and the resulting reconstructed meshes using high resolution and low resolution images. While all pose distribution methods were able to reconstruct the complete mesh using high resolution images, their impact on reconstruction results using low resolution images was much more significant, with the pose distribution method using linear variation with title angle provides the most complete mesh. Table 3 shows the quantification comparison, and the results for low resolution images proves that the linear variation of the title angle give the lowest Hausdorff distance values, while the slight differences in high resolution images were likely caused by slight misalignment when comparing to the ground truth.

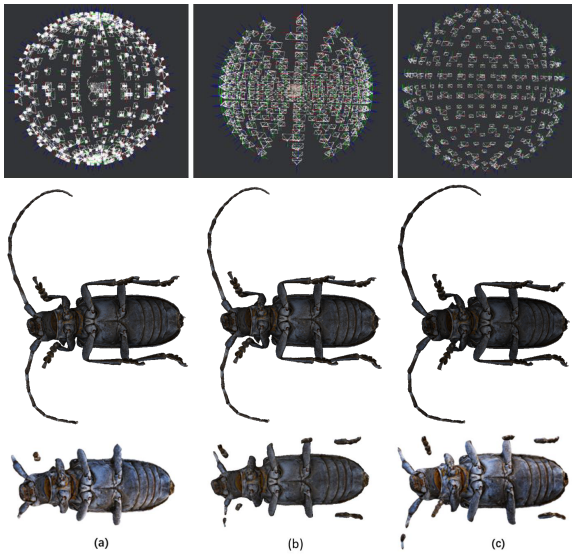


Figure 5: Effect of different pose distributions for image capture: a) constant pan-tilt rotation steps (217 poses), b) linear variation along z axis (189 poses), and c) linear variation with tilt angle (217 poses). From top row to bottom row: camera control diagram, estimated camera poses, corresponding MVS reconstructed meshes using 4320×2880 pixel and 1920×1080 pixel images.

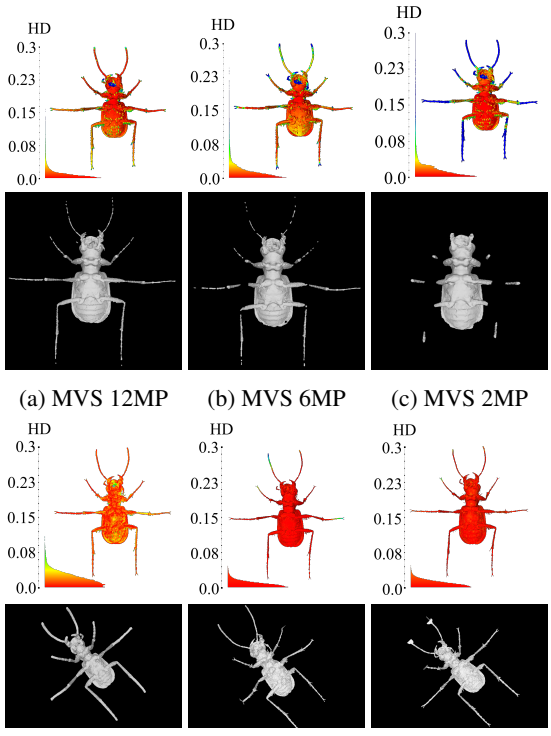
Pose distributions	**HD-12MP**	**HD-2MP**
Constant pan-tilt rot. steps	0.001025	0.02628
Linear variation along z axis	0.000702	0.01734
Linear variation w. tilt angle	0.000821	0.01652

Table 3: Normalised Hausdorff distance (HD) by bounding box diagonal for reconstruction results using different pose distributions for two image resolutions 4320×2880 pixels and 1920×1080 pixels.

3.4. Effect of image resolutions

We performed the reconstruction with MVS, NeRF and NSVF using 4 different image resolutions: 4320×2880, 3024×2016, 1920×1080 and 960×540 pixels. However, in order to speed up the training convergence of NeRF and NSVF, the input images were cropped to a square size, from 1920×1080 pixels to 1080×1080 pixels, and from 960×540 pixels to 540×540 pixels. As shown by Fig. 6 and Table 4, MVS was able to achieve high quality reconstruction results using high resolution images, but its accuracy significantly decreases when using low resolution images, since the camera pose estimation significantly depends on the number of interest points that are extracted and matched.

Unlike MVS, NeRF and NSVF both were able to reconstruct high quality mesh even with low resolution images. In particular NSVF was able to achieve better accuracy using low resolution images than the high resolution result for MVS. Furthermore, it was shown that NSVF is not sensitive to image resolution variation as its results using $2\,\mathrm{Mpx}$ and $0.5\,\mathrm{Mpx}$ images are approximately the same.



(a) MVS 12MP (b) MVS 6MP (c) MVS 2MP

(d) NeRF 2MP (e) NSVF 2MP (f) NSVF 0.5MP

Figure 6: 3D reconstructed meshes of Cicindela Campestris (green tiger beetle) [22, 8] for different image resolutions using ground-truth poses.

For the same set of 334 input images of 1920×1080 pixels, MSV took from 1 to 2 hours to run, NeRF from 8 to 12 hours, NSVF from 20 to 24 hours on the same on Nvidia P100 GPU. NSVF needs at least 16 GB of GPU RAM to perform the training while NeRF and MVS need much less RAM.

Reconst. method	Image resolution	Normalised HD
MVS [11]	4320×2880 (12MP)	0.00197
	3024×2016 (6MP)	0.00293
	1920×1080 (2MP)	0.01220
	960×540 (0.5MP)	NA
NeRF [15]	1920×1080 (2MP)	0.00264
	960×540 (0.5MP)	NA
NSVF [14]	1920×1080 (2MP)	0.00159
	960×540 (0.5MP)	0.00139

Table 4: Normalised Hausdorff distance (HD) by bounding box diagonal for different image resolutions using ground-truth poses. Higher resolution for NeRF and NSVF is not possible due to constraints on GPU memory.

3.5. Effect of estimated poses vs ground truth poses

MVS. For this validation, we use a number of objects of different shapes and material from [22, 8] including: Agrilus Anxius (AA), Anoplophora Chinensis (AC), Fagus Sylvatica (FS), and Melitaea britomartis (MB)[2].

Reconstruction using MVS technique relies on the accuracy of the estimated poses to generate an accurate point cloud for meshing. However, in cases where image resolution is low, or where the object itself has a complex structure, pose estimation could be less accurate. Therefore, we compared the reconstruction results obtained using the estimated poses and the ground truth poses with 4 different specimens.

As shown by Fig. 7 and Table 5, for most specimens, pose estimation was accurate enough to reconstruct a sufficiently accurate meshes, and using ground truth poses leads to a small improvement. However, in the case of the butterfly, as shown by the last column of Fig. 8, where pose estimation was inaccurate as Meshroom to match correct image features the thin symmetric wings. Providing ground truth poses fixes the problem and produce highly accurate reconstructed mesh.

NeRF and NSVF. NeRF with estimated posed failed to generate a 3D model. NSVF however works to some extent and produced a 3D mesh shown in Fig. 9. Compared to NSVF reconstructed mesh using ground-truth poses shown in Fig. 6(e), NSVF reconstructed mesh using estimated posed is quite distorted leading to more than 4 times larger Hausdorff distance.

[2]The 3D models are available at https://skfb.ly/o7rRX

Specimen	Normalised HD	
	Estimated poses	Ground truth pose
AA	0.00154	0.00173
AC	0.00256	0.00201
BF	0.00338	0.00315
MB	NA	0.00144

Table 5: Normalised Hausdorff distance (HD) for MVS reconstruction results using estimated poses versus ground truth poses.

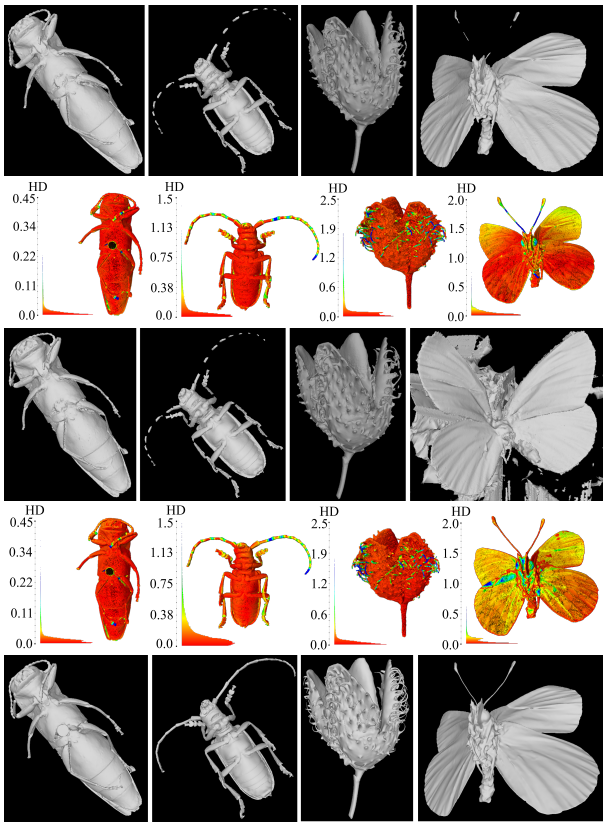
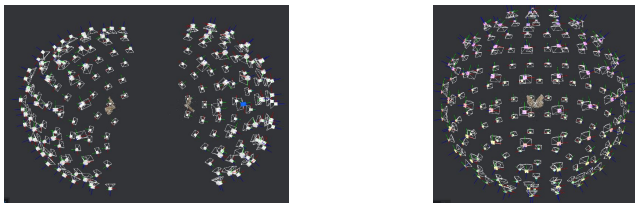


Figure 7: MVS reconstruction results using 4320×2880 pixel images with ground truth poses (top 2 rows), and estimated poses (mid 2 rows) as compared to ground truth meshes (bottom row).

3.6. Effect of different types of object geometries

Reconstruction results using ground truth poses by MVS, NeRF and NSVF are shown in Fig. 10. Accuracy comparison in terms of Hausdorff distance is given by Table 6. Despite using image resolution of 1920 × 1080 pixels, NeRF and NSVF produce 3D reconstructed meshes equivalent or better than MVS using image resolution of 4320×2880 pixels. MVS reconstructed meshes tend to have broken thin parts including legs, wings and spikes. NeRF



(a) Estimated poses (b) G.truth poses

Figure 8: Estimated poses versus ground truth poses for 3D reconstruction of butterfly.

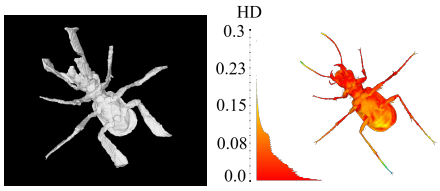


Figure 9: NSVF reconstruction using estimated poses with image resolution of 1920 × 1080 pixels. Normalise Hausdorff is 0.006963.

reconstructed meshes are complete, but they are thicker than ground truth meshes (bottom row of Fig. 7) and lose details. NSVF reconstructed meshes are closest to ground truth meshes, although the butterfly has incomplete antennae.

Specimen	Normalised HD		
	MVS	NeRF	NSVF
AA	0.00173	0.00363	0.00175
AC	0.00201	0.00295	0.00099
FS	0.00315	0.00767	0.00225
MB	0.00144	0.00309	0.00186
Dragonfly	0.01790	NA	0.007963

Table 6: Normalised Hausdorff distance (HD) for different reconstruction methods using ground truth poses.

3.7. Transparent structure and reflective surface

A set of 334 images of 1080×1080 pixels were rendered at uniform poses around the dragonfly model from [16]. Again, Table 6 shows that NSVF produces a more accurate reconstructed mesh than MVS, although NSVF has some errors near the edges of the wings.

For more validation, the Supplementary material also includes the NeRF and NSVF reconstructed results of a spaceship 3D model with transparent cockpit dome and reflective metal surface [23]. Particularly, the surface of glass cockpit and its visible internal structure are successfully reconstructed.

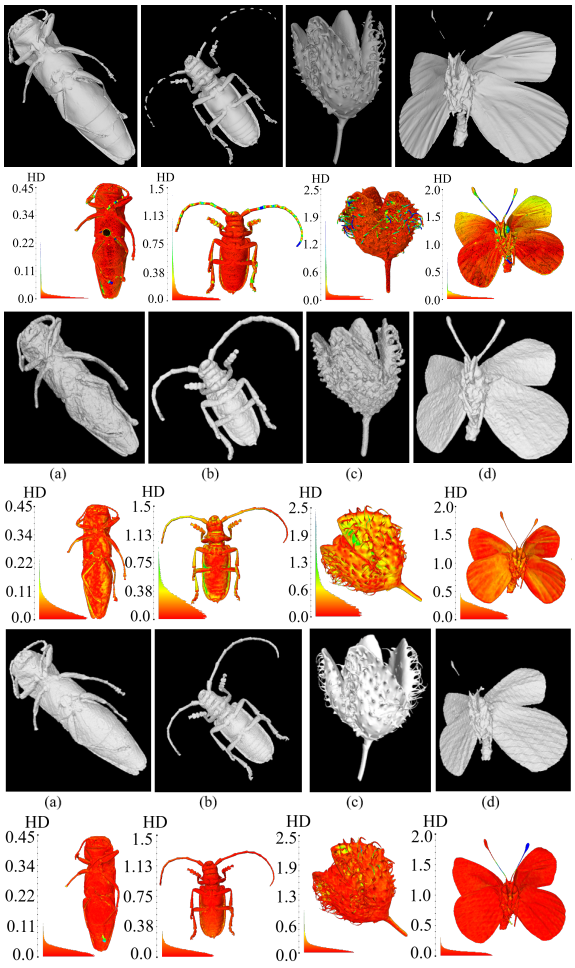


Figure 10: Ground-truth-poses reconstruction results of fine structure dataset by MVS (top rows), NeRF (middle rows) and NSVF (bottom rows). Image resolution for MVS is 4320×2880 pixels, for NeRF and NSVF 1080×1080 pixels.

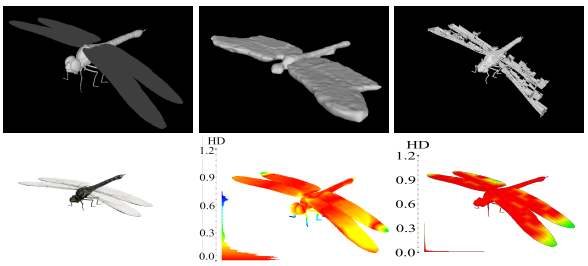


Figure 11: Reconstructed results of dragonfly. Left: ground-truth mesh. Middle: MVS. Right: NSVF.

4. Conclusion

Our study has led to several important practical guidelines for image-based 3D reconstruction of small objects such as insects. To deal with out-of-focus effect, homography transformation for image registration based on lens equation leads to accurate focus-stacked images that allows precise 3D reconstructions. If the lens focal length and magnification are unknown and therefore cannot use the lens equation, the next choice is using a 3D calibration target. Focus stacking without using either of the methods [13, 6] is not recommended, as this likely leads to incomplete and/or inaccurate 3D reconstructed meshes.

Capturing images with a uniform pose distribution around the object results in more complete and accurate 3D reconstructed meshes than the conventional constant rotation step around pan and tilt axis. Relatively uniform pose distribution could be achieved by adjusting the number of rotation steps linearly with tilt angle.

Pose estimated by structure from motion is significantly affected by low image resolutions and objects with flat geometry, reflective surfaces and transparent structures. Using pre-calibrated poses consistently improves the results and the impact can be dramatic for challenging objects, as shown with the butterfly dataset. As a result, using a good motorised stage to obtain ground truth poses is highly recommended.

While high image resolution (above $5\,\mathrm{Mpx}$) is crucial for MVS to produce good pose and depth estimations for an accurate 3D model, image resolution could be as low as $0.5\,\mathrm{Mpx}$ for NSVF to still produce an accurate 3D model as long as camera poses are accurate enough. This suggests that a low-cost 3D reconstruction solution is achievable using NSVF using a good pan-tilt-rail stage (for prior calibrated poses) with a low resolution camera and low-cost macro lens.

The higher resolution meshes are reconstructed by NSVF, however this approach loses very thin structures. On the other hand, NeRF does not achieve the same accuracy but will maintain the thin structures while making them thicker than the ground truth. Even if MVS cannot achieve the quality reached by deep learning algorithms on low-resolution images, it is the only one capable of doing a reconstruction from images at very high resolution. Generally, NSVF and NERF require, respectively, ten and five times more computation and memory resources than MVS.

As the study focuses on using ray-tracing rendered images, it lacks some realistic conditions like image noise, non-uniform background, non-uniform illumination, lens distortion, and misalignment of camera sensor. However, these conditions variations should not change the findings of this study.

Acknowledgment

Chuong Nguyen would like to acknowledge the support of CSIRO Julius Career Award.

References

[1] 3Dflow. 3DF Zephyry - photogrametry software - 3d models from photos. https://www.3dflow.net/3df-zephyr-photogrammetry-software/, 2021. 1

[2] 3DSOM. 3DSOM - 3d models from photos. http://www.3dsom.com/, 2019. 1

[3] Agisoft. Agisoft - Metashape (Photoscan). https://www.agisoft.com/, 2019. 1, 2

[4] AliceVision. ScanRig: Multi-Cameras/Lighting Acquisition Setup for Photogrammetry. https://github.com/alicevision/ScanRig, 2019. 4

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[6] Shah Ariful Hoque Chowdhury, Chuong Nguyen, Hengjia Li, and Richard Hartley. Fixed-lens camera setup and calibrated image registration for multifocus multiview 3d reconstruction. *Neural Computing and Applications*, pages 1–20, 2021. 2, 3, 8

[7] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*, 2008. 3

[8] disc3D. Digital archive of natural history (DiNArDa). https://sketchfab.com/disc3d, 2021. 3, 6

[9] G.D. Evangelidis and E.Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, Oct. 2008. 4

[10] J.M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 368–381. Springer Berlin Heidelberg, 2010. 1

[11] C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, Y. Lanthony, and G. De Lillo. AliceVision Meshroom: An open-source 3D reconstruction pipeline. In *Proceedings of the ACM MUltimedia Systems Conference (MMSys 21)*, 2021. 1, 3, 4, 6

[12] Jared Heinly, Johannes L. Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world in six days. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. 1

[13] Hengjia Li and Chuong Nguyen. Perspective-consistent multifocus multiview 3D reconstruction of small objects. In *Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2019. 2, 8

[14] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2, 3, 6

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 2, 3, 6

[16] mrjeong. Bend swap — dragonfly. https://blendswap.com/blend/14777, 2021. 7

[17] Chuong V Nguyen, David R Lovell, Matt Adcock, and John La Salle. Capturing natural-colour 3D models of insects for species discovery and diagnostics. *PloS one*, 9(4):e94346, 2014. 1, 2

[18] Fabian Plum and David Labonte. scAnt – an open-source platform for the creation of 3d models of arthropods (and other small objects). *PeerJ*, 9:e11155, 2021. 1, 2

[19] RealityCapture. Realitycapture: Mapping and 3d modelling photogrammetry. https://www.capturingreality.com/, 2021. 1

[20] Johannes Lutz Schönberger and contributors. COLMAP: a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline, 2020. [Online; accessed 15-November-2020]. 1

[21] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 1

[22] Bernhard Ströbel, Sebastian Schmelzle, Nico Blüthgen, and Michael Heethoff. An automated device for the digitization and 3D modelling of insects, combining extended-depth-of-field and all-side multi-view imaging. *ZooKeys*, 759(759):1, 2018. 1, 2, 3, 4, 6

[23] thecali. Bend swap — 4060.b spaceship. https://blendswap.com/blend/13489, 2021. 7

[24] Wencheng Wang and Faliang Chang. A multi-focus image fusion method based on laplacian pyramid. *J. Comput.*, 6(12):2559–2566, 2011. 2

[25] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3