Best Foot Forward: Robust Foot Reconstruction in-the-wild

Kyle Fogarty
University of Cambridge
ktf25@cam.ac.uk

Jing Yang
University of Cambridge
jy496@cam.ac.uk

Chayan Kumar Patodi
Hike Medical
chayan@hikemedical.com

Jack Foster
Hike Medical
jack.foster@hikemedical.com

Aadi Bhanti
Hike Medical
aadi@hikemedical.com

Steven Chacko
Hike Medical
steven@hikemedical.com

Cengiz Öztireli
University of Cambridge
aco41@cam.ac.uk

Ujwal Bonde
Hike Medical
ujwal@hikemedical.com

Abstract

High-fidelity 3D foot reconstruction is crucial for prescription orthotics but is hindered by expensive, specialized equipment that limits patient access. We overcome this barrier with the first end-to-end pipeline to reconstruct clinically-accurate foot meshes from simple, self-captured smartphone videos. Our method uniquely solves the core challenges of in-the-wild scanning: we resolve pose ambiguities using SE(3) canonicalization with viewpoint prediction, and then complete partial geometry using an attention-based network. Clinical validation demonstrates that our reconstructions achieve state-of-the-art accuracy and meet prescription-readiness standards, preserving the anatomical fidelity essential for medical intervention. By democratizing high-quality foot assessment, our work unlocks new opportunities for accessible telemedicine, preventative diabetic care, and personalized orthotic treatment. We will release our dataset online at: https://bestfootforward.netlify.app.

1. Introduction

High-fidelity digital twins of the human body represent a foundational technology in autonomous healthcare, with the capacity to deliver significant advancements in personalized medicine. Toward this goal, we present a novel method for high-quality foot geometry capture, tailored to the design of custom foot orthotics. The proposed vision-based workflow automates the acquisition of complex foot geometries, demonstrating robustness against incomplete capture, to enable a fully digital manufacturing process consistent with the tenets of personalized medicine [18]. The methodology has direct implications for a spectrum of applications, including custom prosthetics, preoperative planning, and immersive digital environments such as virtual try-ons and gaming.

While significant progress has been made in reconstructing human anatomy from images [3, 11, 14, 21], creating accurate models of deformable and self-occluding parts like the human foot remains a formidable challenge. Self-captured scanning with commodity devices remains an unsolved problem while traditional approaches, including laser scanners [34], structured-light booths [7], marker-based gait laboratories [26], and impression boxes [19], rely on costly hardware, controlled environments, and skilled operators. Although recent efforts have sought to democratize high-quality foot reconstruction via smartphone video [5, 17], these approaches often depend on LIDAR-equipped devices, limiting accessibility, or they are not robust to in-the-wild partial scans, leading to unreliable geometry. To address this gap, we present the first method that *robustly* reconstructs high-quality foot geometry from a single, self-captured video taken with a standard smartphone.

While classical 3D reconstruction pipelines based on Structure-from-Motion (SfM) and Multi-View Stereo (MVS) excel under controlled conditions [8], they are notoriously brittle for in-the-wild medical capture. The self-scanning of a foot is a particularly challenging instance of this problem; user-captured sequences are often sparse and incomplete due to limited mobility (Fig. 1). Without a strong geometric prior, the resulting reconstructions are not suitable for designing effective orthotics. To overcome

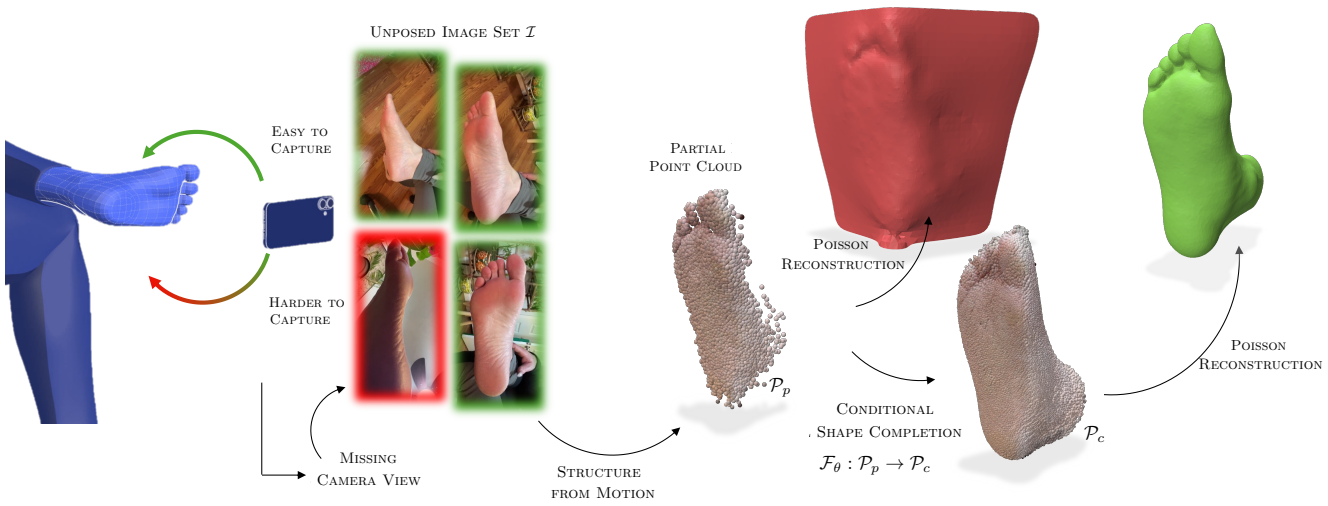

Figure 1. Challenges in foot self-scanning for individuals with reduced mobility: The image highlights the difficulty of capturing the complete foot geometry, especially the underside (red regions), which is harder to access; this limitation often leads to incomplete foot geometry.

this fundamental limitation, we introduce a learning-based framework that robustly infers complete, high-fidelity foot geometry from imperfect data. Our primary contributions are:

1. A novel end-to-end pipeline for foot capture and completion, capable of accurately inferring the occluded geometry from partial and noisy point clouds.
2. The introduction of a comprehensive dataset of 3D foot scans (`Hike3D`), featuring broader demographic diversity than existing public datasets.
3. Extensive experiments demonstrating that our method significantly outperforms both classical pipelines (COLMAP) and state-of-the-art neural rendering techniques (e.g., 3D Gaussian Splatting) in reconstruction robustness, geometric accuracy, and orthotic design suitability.

2. Related Work

Multi-view reconstruction: Recovering 3D geometry from multiple images fundamentally depends on accurate estimation of camera intrinsics and relative poses. These parameters can be acquired through onboard sensors such as Inertial Measurement Units (IMUs) or estimated via sparse reconstruction techniques like Structure-from-Motion (SfM) [32]. Once the camera parameters are established, dense reconstruction is typically performed using Multi-View Stereo (MVS) [33], which computes consistent depth and normal maps across multiple views to produce an oriented point cloud. This point cloud is then converted into a surface mesh using a reconstruction method such as Poisson Surface Reconstruction [13]. A widely used implementation of this full pipeline is COLMAP [32, 33]. Recent advances have explored augmenting traditional multi-view reconstruction with learned priors, leveraging deep neural networks to improve accuracy and robustness. While these methods often incur higher computational costs and longer training times, they have demonstrated improved performance in challenging reconstruction scenarios. Neural rendering methods, such as Neural Radiance Fields (NeRF) [25] and 3D Gaussian Splatting [15], have also become dominant. While visually impressive, current approaches are typically slow to train, require many input views, and lack strong geometric priors, limiting their robustness for accurate 3D reconstruction. In our pipeline, we adopt MVSFormer++ [6], a transformer-based architecture designed to enhance multi-view stereo estimation through learned attention.

Pointcloud Shape Completion: Incomplete or sparse point clouds are a common limitation in 3D reconstruction pipelines, often resulting from occlusions, limited viewpoints, or sensor noise [30, 37]. Shape completion methods aim to infer missing geometry and recover plausible object or scene structures from partial observations [41]. Traditional approaches rely on geometric priors or optimization-based techniques [10, 30], while recent methods leverage deep learning to learn shape priors from large datasets [36, 40]. These learned models have demonstrated strong generalization capabilities, particularly in filling in large missing regions with semantically consistent geometry.

Human Foot Reconstruction: Early attempts in foot modeling relied on Principal Component Analysis (PCA)

[1], but these models offered limited flexibility with restrictive shape spaces. Later approaches employ active sensor technologies, where structured light or depth cameras are used to generate point clouds [20, 22, 39]. However, the point cloud geometries obtained from these sensors are often noisy and incomplete. More recently, Boyne et al. proposed the FIND model [4] leveraging a template deformation strategy guided by an implicit neural network to improve reconstruction accuracy. Similarly, Osman et al. [27] developed SUPR, a PCA-based human foot model designed for seamless integration with the SMPL full-body model [21], enabling expressive and anatomically consistent reconstructions. However, neither FIND nor SUPR are well suited for handling partial or noisy inputs.

3. Problem Setup

We consider a set of unposed images of the foot, denoted as $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, where each image $I_i \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$. Our objective is to reconstruct the complete geometry of the foot. To this end, we define a learnable function $\mathcal{F}_\theta$ that maps the image set $\mathcal{I}$ to a completed point cloud, such that $\mathcal{P}_c = \mathcal{F}_\theta(\mathcal{I})$. To effectively address this, we decompose $\mathcal{F}_\theta$ into two composite functions $\mathcal{F}_\theta := \mathcal{D}_\theta \circ \mathcal{S}$, where $\mathcal{S} : \mathcal{I} \to \mathbb{R}^{Q\times 3}$ generates a dense, yet potentially incomplete, point cloud of the foot from the unposed images, and $\mathcal{D}_\theta : \mathbb{R}^{Q\times 3} \to \mathbb{R}^{M\times 3}$ maps this partial point cloud to the completed point cloud target $\mathcal{P}_c \in \mathbb{R}^{M\times 3}$.

The challenge in learning $\mathcal{F}_\theta$ stems from supervision difficulties across inconsistent vector spaces. The geometric transformations between $\mathcal{D}_\theta$ and $\mathcal{S}$ are in general unknown, creating constraints on pose and scale that complicate end-to-end system development. Our key insight is to address this by decomposing the problem into manageable sub-problems and leveraging synthetic training data at each stage. Then by leveraging a robust canonicalization approach we can effectively connect each stage together. In the following section, we outline our method in more detail.

4. Method

We tackle complete foot reconstruction with a two-phase approach: first, we use Structure-from-Motion (SfM) and Multi-View Stereo (MVS) to estimate camera pose and generate an initial, though potentially incomplete, point cloud. Then, our shape completion module completes missing geometry to produce a dense, complete representation.

A straightforward strategy is to estimate the geometric transformation between the two spaces using Iterative Closest Point (ICP) [2]. However, this often fails because the point clouds produced by SfM/MVS are typically sparse and incomplete, leading to unreliable alignment. To address this, we introduce a viewpoint prediction (VPP) module that robustly estimates the transformation between the SfM/MVS output and the expected input alignment for shape completion.

In the following section, we outline the core components of our reconstruction pipeline, illustrated in Fig. 2. The pipeline begins with two branches: View-Point Prediction (VPP) and SfM & MVS, discussed in Sec. 4.1 and Sec. 4.2, respectively. The VPP module canonicalizes the recovered partial point cloud (Sec. 4.3) before proceeding with foot completion and reconstruction (Sec. 4.4).

4.1. View-point Prediction

The first branch of our architecture, the VPP module, estimates both a bounding box of the foot and the pose relative to a predefined template mesh. Given an unposed image set $\mathcal{I}$ and a reference mesh $\mathcal{M}_{\text{ref}}$, we train a neural network to regress the approximate six degrees of freedom (6-DoF) of the camera pose relative to $\mathcal{M}_{\text{ref}}$. Our method builds on YOLO6D [23], adopting a similar training strategy and leveraging synthetic data; implementation details are in Sec. 5. We represent the VPP module as $\mathcal{V}_\phi$ and define its output for a given image $\mathcal{I}_i \in \mathcal{I}$ as: $(\hat{\mathcal{C}}_i, B_i) = \mathcal{V}_\phi(\mathcal{I}_i)$, where $\hat{\mathcal{C}}_i$ denotes the estimated camera parameters, and B_i represents the bounding box of the foot in the image.

4.2. SfM & MVS

We use a standard structure-from-motion (SfM) pipeline to estimate 3D structure by matching keypoints across views and jointly refining camera poses and a sparse point cloud via bundle adjustment. Specifically, we utilize GLOMAP [29], which from our experiments, observed to give significantly more efficient and scalable global reconstruction compared to COLMAP [32, 33]. For the image-set $\mathcal{I}$, we model the SfM process as:

$$\underbrace{\{\mathcal{C}_1, \cdots, \mathcal{C}_N\}}_{\boldsymbol{\mathcal{C}}} = \mathrm{SfM}(\mathcal{I}),$$

where each $\mathcal{C}_i \in \boldsymbol{\mathcal{C}}$ represents the estimated camera parameters of image $\mathcal{I}_i$. Using the bounding box B_i from the VPP module, we generate segmentation masks via the Segment Anything Model 2 (SAM2) [31] in a zero-shot setting. Denoting the set of all bounding box centers as $\mathcal{B}$, we model the segmentation process as:

$$\underbrace{\{\hat{\mathcal{I}}_1, \cdots, \hat{\mathcal{I}}_N\}}_{\hat{\boldsymbol{\mathcal{I}}}} = \mathrm{SAM}(\mathcal{I}, \mathcal{B}),$$

where $\hat{\mathcal{I}}_1$ is the i-th masked image. To reconstruct a dense point cloud, we employ a multi-view stereo (MVS)

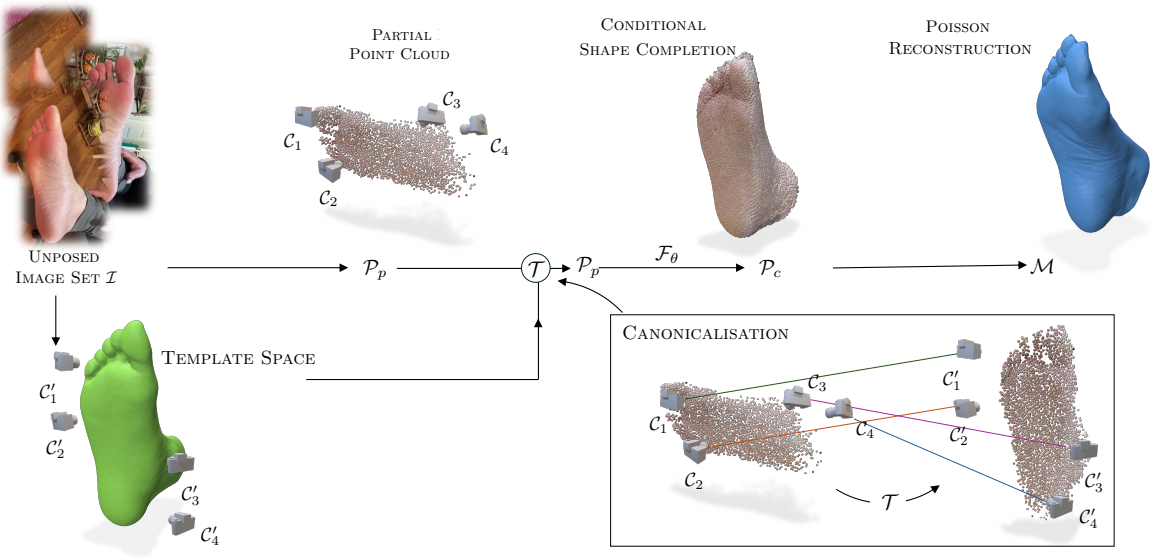

Figure 2. An overview of our reconstruction pipeline: Given an unposed image set $\mathcal{I}$, our method runs two parallel branches: SfM for camera calibration and point cloud generation, and pose estimation. We then canonicalize the point cloud with $\mathcal{T} \in \mathrm{SE}(3)$, apply shape completion $\mathcal{F}_\theta$, and use Poisson reconstruction for the final mesh.

approach [9], which estimates depth by matching pixel correspondences across multiple views and refining depth maps; in this work we leverage the state-of-the-art MVS-Former++ [6] to recover a high-quality point-cloud. Using the camera parameters from GLOMAP and segmentation masks from SAM2, we reconstruct the visible foot geometry pointcloud as:

$$\mathcal{P}_p = \mathrm{MVS}(\mathcal{I}, \mathcal{C}, \hat{\mathcal{I}}).$$

4.3. Point Cloud Canonicalization

Partial point-clouds recovered from image sets have arbitrary poses and scales, complicating their use in downstream shape completion. To address this, we transform the point clouds into a known canonical frame using the camera parameters $\hat{C}_i$ estimated by the VPP module and the depth maps estimated in the MVS process $\mathcal{D}_i$. We present our canonicalization approach in more detail in Algorithm. 1.

4.4. Point Cloud Completion

The second stage of our robust reconstruction pipeline focuses on completing the foot geometry using the learned function $\mathcal{D}_\theta(\mathcal{P})$. For this, we propose an attention-based point cloud completion framework that operates on partially reconstructed foot geometries from the SfM/MVS stage.

Building on recent attention-based approaches to point cloud modeling [35], we formulate completion as an auto-encoding problem, where the model predicts a global latent representation to guide the reconstruction. Our attention mechanism aggregates information across the entire point cloud, capturing both local details and global structural patterns without relying on predefined neighborhood structures. We adopt a coarse-to-fine reconstruction strategy with a scaffold-based skip connection that directly integrates a subset of the input point cloud into the reconstruction process. This scaffold helps maintain fidelity to the observed geometry while enabling the model to infer missing regions effectively and reduce scan noise.

In our standard pipeline, we then employ the screened Poisson surface reconstruction (SPSR) algorithm [12] to generate a mesh, with normals estimated via PCA.

5. Implementation

In this section we provide details of the implementation and training procedure used to construct our reconstruction pipeline.

Datasets: High-fidelity foot geometry datasets are scarce. Foot3D [4] is a valuable resource, but its narrow age and height range prompted us to develop `Hike3D`, a more diverse dataset for orthotics research. The `Hike3D` dataset comprises 15 participants recruited in the United States from a variety of occupational backgrounds. Data acquisition was performed using an EinStar structured-light scanner in a controlled indoor setting. Participants were seated with one foot extended, barefoot, in a neutral posture. Each foot was scanned individually, beginning from the plantar surface and sweeping around to the

Algorithm 1 Canonicalization

Require: Reference mesh: $\mathcal{M}_{\text{ref}}$,
Camera parameters: C_i [SfM], $\hat{C}_i$ [VPP],
Depth maps: D_i, Point-cloud: $\mathcal{P}_p$.

Select k points $p_1, p_2, \ldots, p_k$ from $\mathcal{M}_{\text{ref}}$
for $i \leftarrow 1$ to N **do**
  for $j \leftarrow 1$ to k **do**
    $q_{i,j} \leftarrow \text{Project}(p_j, \hat{C}_i)$
    $d_{i,j} \leftarrow D_i(q_{i,j})$
    $p'_{i,j} \leftarrow \text{BackProject}(q_{i,j}, d_{i,j}, C_i)$
  end for
end for
Compute Centroids c_j for each $j = 1 \ldots k$
Procrustes: $\{R, t, s\} \leftarrow \text{Procrustes}(\{p_j\}, \{c_j\})$
$\mathcal{P}'_p \leftarrow s(R \cdot \mathcal{P}_p + t)$
ICP refinement: $\{R', t'\} \leftarrow \text{ICP}(\mathcal{P}'_p, \mathcal{M}_{\text{ref}})$
$\mathcal{P}_{\text{aligned}} \leftarrow R' \cdot \mathcal{P}'_p + t'$
return $\mathcal{P}_{\text{aligned}}$

dorsal surface, capturing full geometry from bottom to top. Scans were cleaned, aligned to a common reference frame, and manually inspected for quality control, with their geometric accuracy evaluated in an offline assessment. To broaden demographic coverage and strengthen design robustness, we integrate `Hike3D` with Foot3D. We provide an overview of the distribution of `Hike3D` compared with [4] in Fig. 3. We will release the dataset as part of this work.

VPP module: We train the VPP module using synthetically generated images of meshes from our training set. To ensure diversity, we utilize 740 HDR backgrounds, creating various background combinations for our 50k synthetic images. We use Blender to simulate lighting variations and skin tone modulation, enhancing the realism of the synthetic data. We then fine-tune the model using 5k real images. Throughout the process, we apply the same augmentations and loss functions as in the original work [23].

Foot Completion Module: We train the foot completion module using a simulated scanning setup to generate paired partial and complete geometries. To improve robustness against noise and SE(3) perturbations, we apply data augmentations during training. Our dataset combines `Hike3D` and Foot3D, splitting the dataset with 80% for training and 20% for testing. Each mesh was augmented with 10 spatial transformations (shifts, scaling, rotations), followed by five virtual scans per transformation, yielding 2000 training and 250 testing pairs. Supervision is applied by minimizing the Chamfer distance between predicted and ground-truth point clouds at intermediate steps of the network.

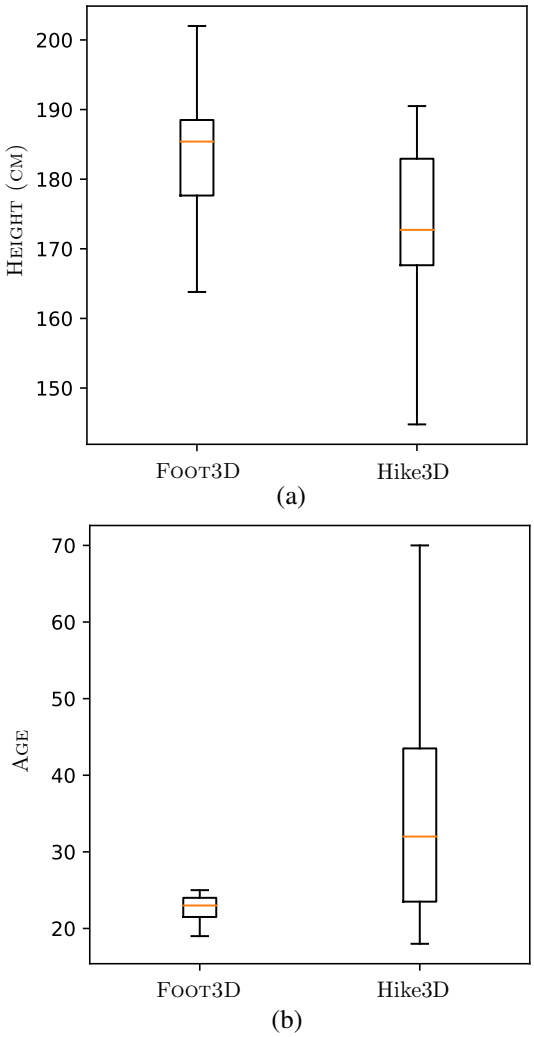

Figure 3. Boxplots summarizing the distribution of heights and ages in our dataset, `Hike3D`, reveal a broader diversity in heights and significantly greater variation in ages compared to prior datasets. This age variation is particularly important, as feet undergo changes and deformations over time due to factors like footwear and activity levels.

6. Experiments

To evaluate robustness, we conduct two experiments: one using paired videos and high-quality 3D scans for quantitative error analysis, and another using video-only data, where clinicians score reconstructions based on visual assessment.

6.1. Experimental Setup

Completion Module Evaluation: To assess the efficacy of our foot completion module, we perform a quantitative comparison against three established foot modeling techniques. Each method is tasked with fitting a model to

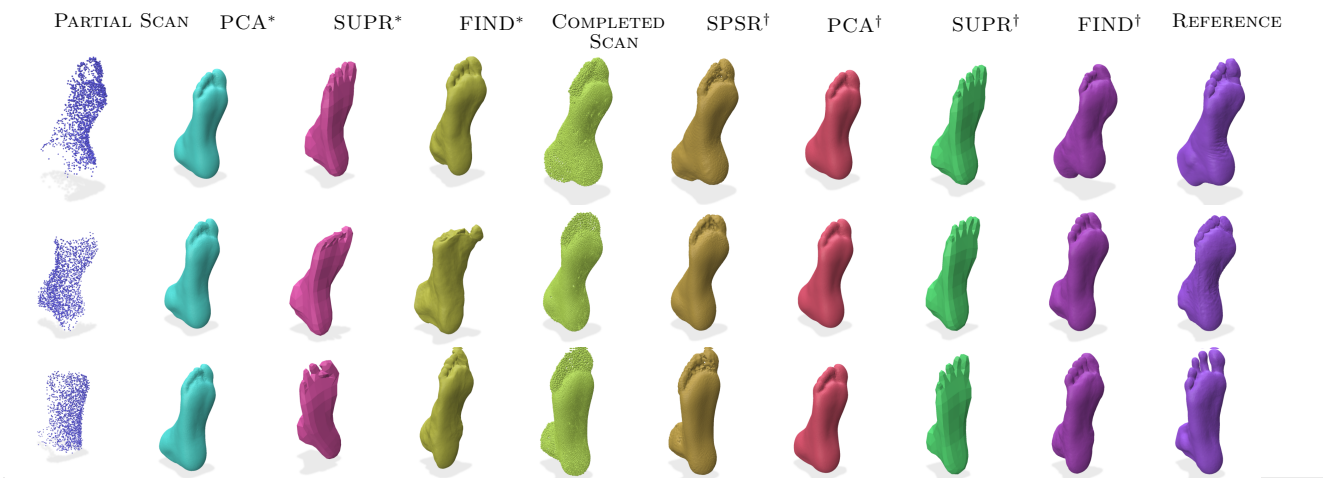

Figure 4. Figure shows partial scan reconstruction results. Methods marked $*$ are optimized on the input scan, while † denotes optimization on our completed point cloud, which recovers geometry much closer to the reference scans.

Method	CD (↓) (10^{-2})	HD (↓) (10^{-2})
PCA	4.46 ± 1.24	12.28 ± 3.26
SUPR	12.74 ± 3.78	34.61 ± 7.91
FIND	15.95 ± 6.11	37.19 ± 14.36
Ours	2.29 ± 0.56	9.51 ± 3.76
SPSR + Ours	2.81 ± 0.77	10.20 ± 3.13
PCA + Ours	3.93 ± 1.08	11.37 ± 3.02
SUPR + Ours	7.08 ± 1.79	27.95 ± 5.43
FIND + Ours	3.46 ± 1.26	10.05 ± 3.50

Table 1. Average Chamfer Distance (CD) and Hausdorff Distance (HD) results. The top section shows results from direct fitting to partial inputs, middle is results for our generated point cloud, and bottom is results from fitting on our generated point clouds. Reported values include ± 1 standard deviation over the test dataset.

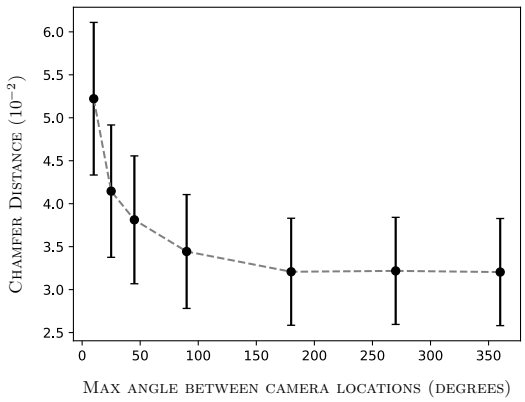

Figure 5. Chamfer Distance between predicted foot mesh and ground truth vs. camera scanning angle.

both the original incomplete scans and our completed point clouds. The baseline methods are: (1) a PCA-based model [1] with vertex correspondences obtained using functional maps [24, 28]; (2) the parametric SUPR foot model [27]; and (3) FIND [4], which utilizes a large latent space for detailed shape and pose control. For all methods, we optimize shape, pose, and transformation parameters via gradient descent with the Adam optimizer [16], minimizing the Chamfer distance. Final reconstruction accuracy is quantified using both Chamfer and Hausdorff distances.

End-to-End Pipeline Evaluation: We evaluate the performance of our entire reconstruction pipeline using video footage captured in uncontrolled, real-world settings. Our method is benchmarked against two prominent 3D reconstruction pipelines: (1) COLMAP [32], a standard Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline that reconstructs dense geometry, and (2) Gaussian Opacity Fields (GOF) [38], a state-of-the-art method for adaptive surface reconstruction from images. Our dataset consists of 30 videos from smartphone videos, featuring diverse subjects, lighting, and backgrounds. For qualitative assessment, three expert clinicians in foot anatomy and orthotics reviewed randomized renderings from all three methods. They rated each reconstruction on a 5-point Likert scale based on three criteria: (1) anatomical accuracy, (2) completeness of the foot structure, and (3) suitability for orthotic design.

6.2. Experimental Results

Completion Module Performance: As detailed in Table 1, our point cloud completion method significantly outperforms the template-based fitting approaches when applied to incomplete data, achieving lower Chamfer and

Hausdorff distances. Furthermore, when the outputs are meshed using Screened Poisson Surface Reconstruction (SPSR), the surfaces generated from our completed point clouds exhibit the lowest Chamfer distance, affirming the benefit of our approach for foot mesh recovery. The qualitative results in Figure 4 visually corroborate these quantitative improvements, showing more plausible and complete foot geometry.

Robustness to Partial Views: We analyzed our shape completion module's performance under varying degrees of data incompleteness. In a simulated scanning environment, we controlled the partialness of the input scan by restricting the maximum viewing angle between the virtual camera and the foot surface. As shown in Figure 5, the reconstruction error (Chamfer Distance) systematically decreases as the viewing angle—and thus the surface coverage—increases. Notably, the error rate plateaus around 90°, a viewing range that is practically achievable in self-scanning scenarios. This demonstrates that our module delivers robust and accurate completions.

End-to-End Clinical Assessment: The results of our end-to-end evaluation are summarized in Figure 6 (a)–(c). The clinical assessment reveals that our method consistently outperforms both baselines across all criteria. Reconstructions from COLMAP were frequently cited for poor anatomical accuracy and major geometric deviations. While GOF performed better, its results were rated as inconsistent. In contrast, our method achieved the highest and most stable scores for anatomical fidelity, completeness, and suitability for orthotics. These findings underscore our system's potential for reliable use in clinical applications, such as the design and manufacturing of custom foot orthotics.

7. Discussion

Our results demonstrate that our end-to-end pipeline significantly improves foot geometry reconstruction, achieving lower Chamfer and Hausdorff distances while maintaining consistency with input data. The foot completion module, leveraging learned priors, successfully reconstructs plausible geometries from sparse data, addressing the limitations of template-based methods, which showed constrained shape variability in our evaluations. Our approach enables robust reconstruction across diverse and incomplete inputs, as reflected in its consistently high surface completeness scores. Furthermore, by integrating completion and canonicalization, our method effectively mitigates occlusions and partial view challenges, leading to more accurate and reliable reconstructions, as evidenced by its superior anatomical fidelity and quality ratings.

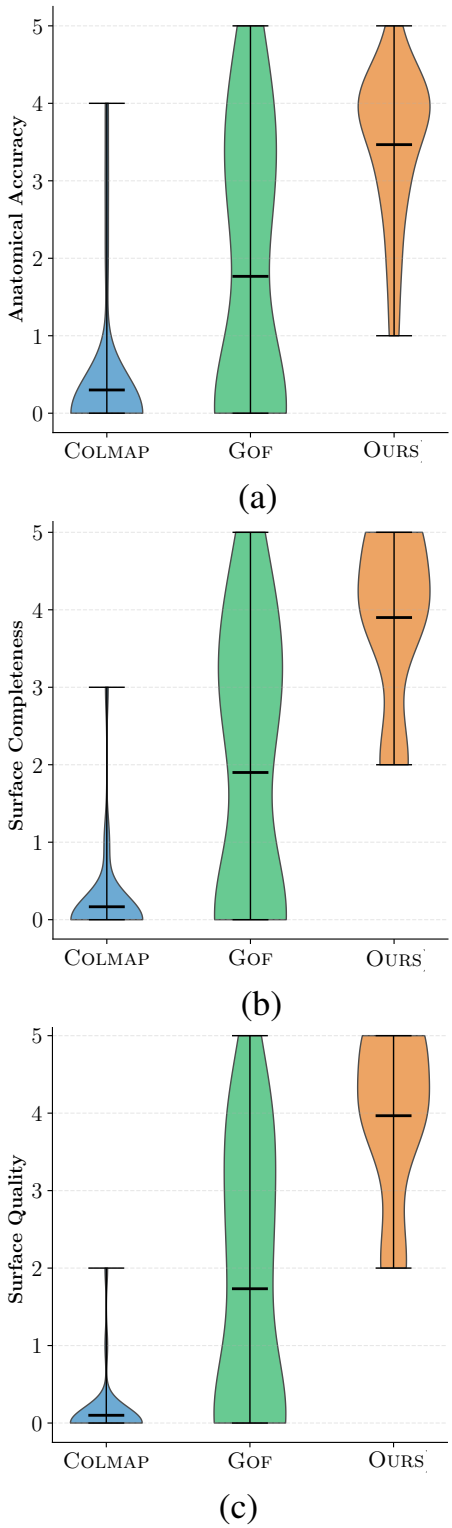

Figure 6. Plots of the distribution of aggregated clinical scores for each methods assessing (a) anatomical accuracy (b) completeness (c) surface quality.

While overall performance is strong, fine structures such as toes can be more challenging to reconstruct accurately, particularly in sequences where the static scene assumption is mildly violated due to subtle motion. Additionally, the current model lacks uncertainty quantification for highly out-of-distribution geometries, which we plan to explore in future work.

8. Conclusion

We introduced a novel end-to-end pipeline for reconstructing foot geometry from self-scanned mobile videos, addressing key limitations of existing methods. Our proposed method provides robust foot reconstruction, even from partial observation. Extensive evaluation demonstrated that our method outperforms baseline approaches, achieving lower Chamfer and Hausdorff distances while preserving consistency with input geometry. These findings underscore the effectiveness and robustness of our approach, particularly for self-scanning applications, paving the way for improved foot reconstruction in real-world settings.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [EP/S023917/1].

References

[1] Edmée Amstutz, Tomoaki Teshima, Makoto Kimura, Masaaki Mochimaru, and Hideo Saito. PCA based 3D shape reconstruction of human foot using multiple viewpoint cameras. In *Computer Vision Systems: 6th International Conference*, 2008. 3, 6

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 3

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 1

[4] Oliver Boyne, James Charles, and Roberto Cipolla. Find: An unsupervised implicit 3d model of articulated human feet. In *British Machine Vision Conference*, 2022. 3, 4, 5, 6

[5] Oliver Boyne, Gwangbin Bae, James Charles, and Roberto Cipolla. Found: Foot Optimisation with Uncertain Normals for surface Deformation using synthetic data. In *Winter Conference on Applications of Computer Vision*, 2024. 1

[6] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. *arXiv preprint arXiv:2401.11673*, 2024. 2, 4

[7] Hansjoerg Gaertner, Jean-Francois Lavoie, Eric Vermette, and Pascal-Simon Houle. Multiple structured light system for the 3d measurement of feet. In *Three-Dimensional Image Capture and Applications II*, pages 104–114. SPIE, 1999. 1

[8] Klaus Häming and Gabriele Peters. The structure-from-motion reconstruction pipeline–a survey with focus on short image sequences. *Kybernetika*, 2010. 1

[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 4

[10] Hui Huang, Dan Li, Hao Zhang, Uri Ascher, and Daniel Cohen-Or. Consolidation of unorganized point clouds for surface reconstruction. *ACM transactions on graphics (TOG)*, 28(5):1–7, 2009. 2

[11] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1

[12] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 4

[13] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 2

[14] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics*, 2023. 1

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[16] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Felix Kok, James Charles, and Roberto Cipolla. Footnet: An efficient convolutional network for multiview 3d foot reconstruction. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[18] Marco Leite, Bruno Soares, Vanessa Lopes, Sara Santos, and Miguel T Silva. Design for personalized medicine in orthotics and prosthetics. *Procedia CIRP*, 2019. 1

[19] AKL Leung, JCY Cheng, and AFT Mak. Orthotic design and foot impression procedures to control foot alignment. *Prosthetics and orthotics international*, 28(3):254–262, 2004. 1

[20] Samuel J Lochner, Jan P Huissoon, and Sanjeev S Bedi. Development of a patient-specific anatomical foot model from structured light scan data. *Computer methods in biomechanics and biomedical engineering*, 2014. 3

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 1, 3

[22] Nolan Lunscher and John Zelek. Point cloud completion of foot shape from a single depth map for fit matching using deep learning view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 3

[23] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-6d-pose: Enhancing yolo for single-stage monocular multi-object 6d pose estimation. In *International Conference on 3D Vision*, 2024. 3, 5

[24] Simone Melzi, Jing Ren, Emanuele Rodola, Abhishek Sharma, Peter Wonka, and Maks Ovsjanikov. Zoomout: Spectral upsampling for efficient shape correspondence. *arXiv preprint arXiv:1904.07865*, 2019. 6

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[26] Matteo Moro, Giorgia Marchesi, Filip Hesse, Francesca Odone, and Maura Casadio. Markerless vs. marker-based gait analysis: A proof of concept study. *Sensors*, 22(5):2011, 2022. 1

[27] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Supr: A sparse unified part-based human representation. In *European Conference on Computer Vision*, 2022. 3, 6

[28] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics*, 2012. 6

[29] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, 2024. 3

[30] Mark Pauly, Niloy J Mitra, Joachim Giesen, Markus H Gross, and Leonidas J Guibas. Example-based 3d scan completion. In *Symposium on geometry processing*, pages 23–32, 2005. 2

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 6

[33] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 3

[34] Rui Silva, Bruna Silva, Cristiana Fernandes, Pedro Morouço, Nuno Alves, and António Veloso. A review on 3d scanners studies for producing customized orthoses. *Sensors*, 24(5): 1373, 2024. 1

[35] Jun Wang, Ying Cui, Dongyan Guo, Junxia Li, Qingshan Liu, and Chunhua Shen. Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on artificial intelligence*, 2024. 4

[36] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 790–799, 2020. 2

[37] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 2

[38] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024. 6

[39] Munan Yuan, Xiaofeng Li, Jinlin Xu, Chaochuan Jia, and Xiru Li. 3d foot scanning using multiple realsense cameras. *Multimedia Tools and Applications*, 2021. 3

[40] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018. 2

[41] Zhiyun Zhuang, Zhiyang Zhi, Ting Han, Yiping Chen, Jun Chen, Cheng Wang, Ming Cheng, Xinchang Zhang, Nannan Qin, and Lingfei Ma. A survey of point cloud completion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:5691–5711, 2024. 2