Fast and Accurate 3D Foot Reconstruction from a Single Image

Joaquin SANCHIZ, Eduardo PARRILLA, Jordi URIEL, Alfredo BALLESTER, Sandra ALEMANY
Instituto de Biomecánica de València, Universitat Politècnica de València, Valencia, Spain

https://doi.org/10.15221/24.13

Abstract

Obtaining accurate 3D reconstructions of the human foot from 2D images holds significant importance in various fields, including anthropometry, footwear design, and medical diagnostics. In this study, we propose a novel neural network-based approach for reconstructing a 3D mesh of the foot from a single image. Our method integrates multiple sources of information, including binary segmentation masks and 2D keypoint estimation. By leveraging Principal Component Analysis (PCA) to represent foot morphology in a low-dimensional space, we infer the parameters needed for 3D mesh reconstruction, including rotation and translation parameters for alignment with the input image. Our approach builds upon recent advancements in deep learning for 3D reconstruction from images and demonstrates promising results in accurately capturing foot morphology. The model has been trained with two datasets: one consisting of 1M synthetic samples and another with 500K augmented real samples. Validation on a test subset of over 674 samples resulted in a PA-MPJPE (Procrustes-Aligned Mean Per Joint Position Error) of 0.9 mm. Furthermore, the real-time capability of our method makes it suitable for applications in augmented reality, such as virtual try-on and improvements in user experience and precision of phone-based foot scanning solutions.

Keywords: anthropometry, footwear design, size recommendation, 3D foot, foot reconstruction, real time

1. Introduction

The human foot, with its intricate structure and complex biomechanics, plays a crucial role in our daily lives. Understanding its morphology is essential in various fields, including medicine, sports science, and footwear design [1], [2], [3], [4], [5]. Accurate 3D modeling of the foot is fundamental for diagnosing and treating foot-related conditions, improving athletic performance, and designing comfortable and ergonomic footwear. Traditionally, acquiring 3D foot models has relied on techniques like 3D scanning, which, while accurate, can be expensive and time-consuming, often requiring specialized equipment and expertise [6], [7].

Subsequently, methodologies based on parametric shape models emerged, enabling the reconstruction of the foot using only a few images [4], [5]. These developments have revolutionized the way we capture and analyze foot morphology, making the process more scalable and accessible to a broader audience, making it possible to find the most suitable footwear model for your feet, the best size or to get personalized insoles, shoes or boots [8], [9], [10], [11], [12].

Recent advancements in computer vision and deep learning have opened doors to explore new efficient and accessible methods for 3D shape reconstruction using a single 2D image [13], [14]. Building on these advancements, this research focuses on developing a novel method capable of generating a 3D mesh of the foot from a single 2D image. Unlike conventional methods that directly regress vertex positions [15], [16], [17], our approach leverages the power of statistical shape modeling [18], [19]. More specifically, the network learns to infer the parameters of a Principal Component Analysis (PCA) [20] model, which represents the 3D foot shape as a linear combination of principal components. This approach allows for a more compact and generalizable representation of the 3D shape, reducing the complexity and improving the robustness of the model. Additionally, the network estimates the rotation and translation parameters, enabling accurate pose estimation within the 3D space. This dual estimation ensures that the reconstructed foot model is not only anatomically accurate but also correctly oriented and positioned in the 3D coordinate system.

The implications of this work are significant. The efficiency of our model enables real-time applications, opening possibilities in augmented reality (AR). This can revolutionize virtual try-ons and interactive fitting systems, providing consumers with instant, accurate visualizations of how footwear will fit and look. Moreover, it has the potential to create better phone-based foot scanning experiences making it faster, simpler and more accurate. Overall, the integration of deep learning techniques with traditional statistical models represents a promising direction for advancing the field of 3D foot reconstruction.

2. Related work

Estimating a 3D human mesh from a single image is a challenging task that has been extensively studied in the field of computer vision and machine learning. The existing methods can be broadly categorized into two groups: parametric and non-parametric approaches.

Parametric methods rely on a predefined 3D human body model, such as the Skinned Multi-Person Linear (SMPL) model [21], to estimate the 3D mesh. These methods typically involve optimizing the model parameters to fit the input image. For example, Kanazawa et al. [22] proposed an end-to-end framework that uses a convolutional neural network (CNN) to predict the SMPL model parameters from a single image. Similarly, Pavlakos et al. [23] used a CNN to estimate the 3D human pose from keypoint estimation and shape from segmentation using a single image, and then fitted the SMPL model parameters to the estimated pose and shape. Other parametric methods have used different 3D human body models, such as the SCAPE model [24] or the Adam model [25]. For instance, Dibra et al. [26] used a CNN to estimate the SCAPE model parameters using silhouettes from heat kernel signature (HKS) [27].

Non-parametric methods, on the other hand, do not rely on a predefined 3D human body model. Instead, they use a data-driven approach to estimate the 3D mesh from the input image. For example, Lin et al. [16] used a CNN and a Graph Convolutional Neural Network to predict a 3D point cloud from a single image, and then used a mesh reconstruction algorithm to obtain the 3D mesh. Similarly, Onizuka et al. [15] used a CNN to predict a 3D occupancy grid from a single image, and then used a marching cubes algorithm to obtain the 3D mesh. In the context of human body mesh estimation from video, You et al. [28] directly regress the mesh vertices using spatial-temporal image features and 3D pose information.

Recent advancements have also explored the use of generative models and implicit neural representations for 3D reconstruction. Neural Radiance Fields (NeRFs), for example, have been employed to generate high-fidelity 3D reconstructions from multi-view images [29]. Additionally, Occupancy Networks [30] and Implicit Differentiable Renderer (IDR) [31] represent surfaces implicitly and have shown promise in reconstructing detailed 3D shapes from sparse data. Despite their potential, these methods require multi-view inputs or volumetric data, which can be limiting in scenarios with only a single view available.

While these approaches have demonstrated remarkable success in 3D human pose and shape estimation, they often focus on the entire body and may not capture the intricate details of the foot. Furthermore, directly regressing vertex positions can be computationally expensive and may lead to inconsistencies in the generated meshes.

3. Method

Our proposed method addresses these limitations by combining the strengths of statistical shape modeling and deep learning. The network architecture follows the typical classic encoder-decoder structure showed in Figure 1. In our design, both the encoder and decoder components have been modularized independently to allow for flexibility and adaptability according to specific requirements. This modularization enables easy interchangeability of different encoder and decoder architectures, catering to diverse needs in terms of performance and accuracy.

The encoder takes a 2D foot image as input and processes it through several layers to extract meaningful features. These features capture the shape, texture, and spatial relationships within the image, providing a compressed representation of the foot. Additionally, the encoder generates a binary segmentation mask that distinguishes the foot region from the background, and a set of anatomical keypoints in the 2D space.

The decoder takes the concatenated feature vector and 2D keypoints as input and maps them onto the final output parameters. These parameters include the PCA coefficients that define the 3D foot shape from our parametric shape model [4], as well as the rotation and translation values that determine the foot's pose in 3D space relative to the camera reference system.

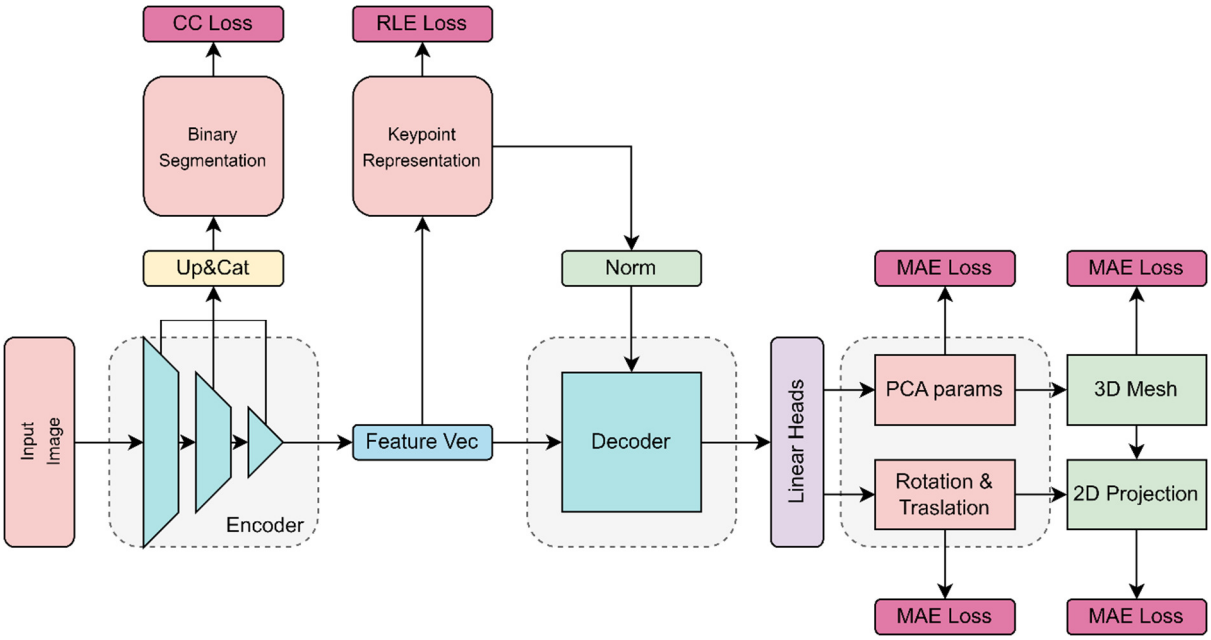

Figure 1: Model architecture

3.1. Intermediate heads

The model is designed with two output heads that facilitate the creation of intermediate representations [32] essential for the final 3D reconstruction of the foot.

Segmentation output: This head generates a binary mask that highlights the foot within the input image. By distinguishing the foot from the background, the mask provides a clear and detailed outline of the foot's shape. This segmentation is crucial for isolating the foot and ensuring accurate subsequent processing steps.

Keypoints output: This head projects nine keypoints through a dense layer [33], locating critical anatomical landmarks of the foot. These keypoints include important regions such as the heel, the arch, and the tips of the toes. The accurate identification of these points is vital for capturing the geometric features of the foot, which are essential for constructing a precise 3D model.

3.2. Main heads

The decoder processes the encoded information, integrating the keypoints output to derive the final parameters necessary for 3D model reconstruction.

PCA output: These parameters define the principal components that describe the foot's shape variation. By reducing the dimensionality of the data, they capture the most significant features necessary for accurately modeling the foot's geometry.

Rotation output: These parameters specify the orientation of the foot. They are crucial for positioning the 3D model correctly relative to the input image, ensuring that the reconstructed model aligns accurately with the original perspective.

Translation output: These parameters determine the foot's position within the 3D coordinate system. They are essential for translating the foot to its correct spatial location, providing an accurate representation of its placement in the reconstructed model.

The model reconstructs a detailed 3D mesh of the foot using PCA parameters and the mean shape. This mesh captures the foot's unique geometry with approximately 5,000 vertices. Following reconstruction, rotation and translation parameters are applied to align and position the mesh accurately within the 2D pixel space, ensuring it reflects the correct orientation and location observed in the input image.

4. Training

The data that our model learns from determines how effective it is. The training dataset is the foundation of our model's power in the context of 3D modeling from 2D photos. This section explores the details of our training data, including its origin, preparation, and model training.

4.1 Data

Our training process utilized two distinct datasets: a set of 500K augmented real-world samples and a collection of 1M synthetic samples. These datasets collectively form the basis for training our deep learning model to generate accurate 3D models from 2D images. Every training sample is composed by: a 2D image and a camera matrix K as input; and a set of keypoint coordinates, binary mask, 2D projection points, PCA shape parameters, rotation parameters and translation parameters as ground truth.

Augmented Avatar 3D Feet dataset: This dataset consists of real-world images taken from 3 different points of view (one upper view and two lateral views) as shown in Figure 2 that have been augmented to enhance diversity and variability [34]. Augmentation techniques include rotation, scaling, and color adjustments and background replacement, ensuring the model learns from a broad spectrum of real-world scenarios. These samples have been selected from the Avatar 3D Feet dataset, in a certified set of high quality and minimal error data [5].

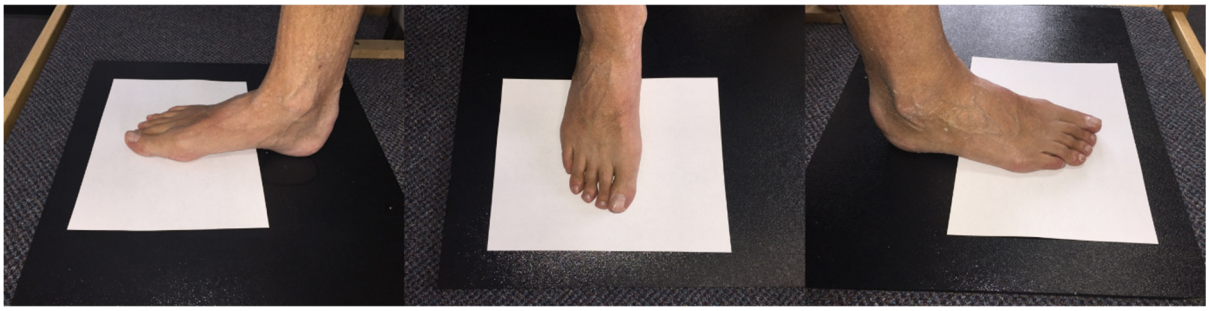

Figure 2: Real sample

Synthetic Samples: Comprising entirely synthetic images generated using computer graphics techniques, this dataset provides additional training examples that cover a wide range of simulated conditions. Synthetic data allows the model to learn robustly across various scenarios that may not be represented adequately in real-world data alone [35], [36], [37]. This meshes have been generated by random PCA parameters following 3 times the standard deviation, then covered with a random texture (selected from a set of 1500 textures extracted from the Avatar 3D Feet dataset) and projected into a random background as you can see in Figure 3. Using a virtual camera, over 20 points of view have been placed in order to represent a complete set of possibilities.

Figure 3: Synthetic sample. On the left, the generated mesh with the texture. On the right, the generated mesh projected on a random background.

4.2 Training process

The model has been designed and trained on Pytorch Lightning for 50 epochs using Adam [38] optimizer, learning rate of 1e-4 with cosine decay, batch size of 36 due to resource limitations and 3 GPUs. Training process takes about 4 days to complete.

5. Metrics

5.1. Evaluation metrics during training

Mean Vertex Error: this metric is used to evaluate the accuracy of 3D reconstructions by measuring the average distance between corresponding vertices of the predicted and ground truth 3D meshes. In this context, MVE functions similarly to the Mean Absolute Error (MAE) but is specifically adapted for 3D coordinates.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{n=1}^{\mathrm{N}}\left(|\widehat{\mathrm{y}_n^{\mathrm{x}}} - \mathrm{y}_n^{\mathrm{x}}| + \left|\widehat{\mathrm{y}_n^{\mathrm{y}}} - \mathrm{y}_n^{\mathrm{y}}\right| + |\widehat{\mathrm{y}_n^{\mathrm{z}}} - \mathrm{y}_n^{\mathrm{z}}|\right)$$

where $\mathrm{y}_n^{\mathrm{x}}$, $\mathrm{y}_n^{\mathrm{y}}$ and $\mathrm{y}_n^{\mathrm{z}}$ are the first, second and third components of the 3D predicted vertex *n* and $\widehat{\mathrm{y}_n^{\mathrm{x}}}$, $\widehat{\mathrm{y}_n^{\mathrm{y}}}$and $\widehat{\mathrm{y}_n^{\mathrm{z}}}$ are the first, second and third components of the 3D ground truth vertex *n.*

Projection Error: this metric quantifies the discrepancy in pixels between the projected 3D model points and their corresponding 2D image points. It measures how accurately the 3D points, when projected onto a 2D plane using a perspective transformation, match the actual 2D coordinates in the image. Lower values indicate a more precise alignment of the 3D model with the image.

Keypoint Accuracy: is a metric used to evaluate the precision of estimated keypoints in relation to their true positions (ground truth). It is calculated as the percentage of correctly predicted keypoints that fall within a predefined threshold distance from the ground truth keypoints. This threshold is usually specified in pixels and defines the maximum allowable deviation for a prediction to be considered accurate. If the estimated keypoint deviation from the ground truth is less than or equal to this threshold, the prediction is counted as correct.

Segmentation Accuracy: The Dice coefficient [39] for binary segmentation is the metric chosen for evaluating the overlap between the predicted binary segmentation mask and the ground truth mask. It is calculated as twice the area of overlap divided by the total number of pixels in both the predicted and ground truth masks. This metric ranges from 0 to 1, with higher values indicating a better match and thus more accurate segmentation.

5.2. Evaluation metrics during testing

The key metric used for evaluation was the Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) [32]. This metric measures the average Euclidean distance between the predicted 3D vertex positions and the ground truth, after performing a Procrustes analysis to align the predicted pose with the ground truth, effectively removing the effects of translation, rotation, and scale.

6. Results

6.1. Training results

For the evaluation of the training of this model we have obtained the following metrics over a 500K samples test dataset, shown in Table 1:

Metric	Score
Mean Vertex Error	1.9 mm
Projection Error	12 px
Segmentation Accuracy	0.99
Keypoint Accuracy	80%

Table 1: Metric scores

6.2. Quantitative evaluation

To quantitatively assess the accuracy of our 3D foot reconstruction model from 2D images, we utilized a subset of 674 samples from the 3D Avatar Feet [4] dataset.
Our model achieved a PA-MPJPE of 0.9 millimeters on this dataset, indicating a high level of accuracy in capturing the detailed structure of the foot. A detailed heatmap of the vertex error is shown in *Figure 4: Error map in millimeters*Figure 4 showing how the mean vertex error is distributed over the foot surface.

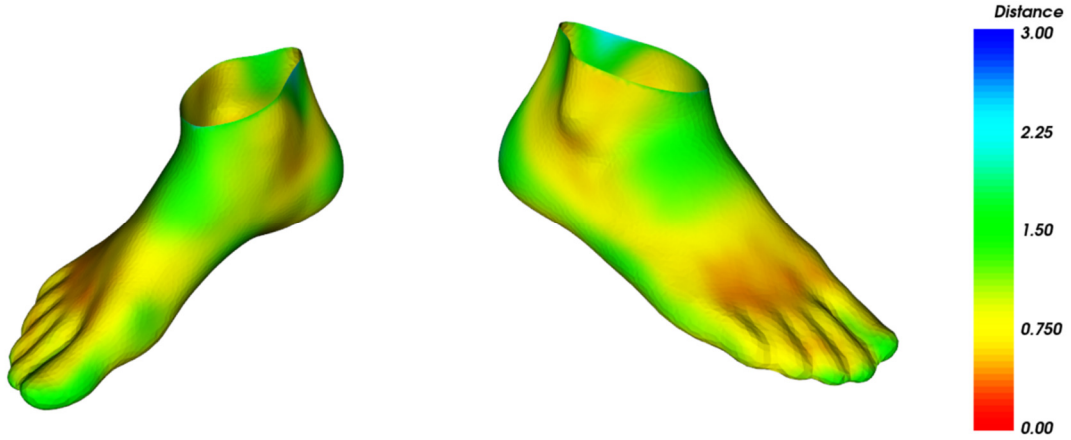

Figure 4: Error map in millimeters

6.3. Qualitative evaluation

The visual evaluation of the process involved observing the reconstructions projected onto various images and videos. Figure 5 presents several examples that illustrate the results of the reconstruction process. This detailed analysis aimed to verify the accurate estimation of the shape, position, and orientation of the reconstructed foot on the input image.

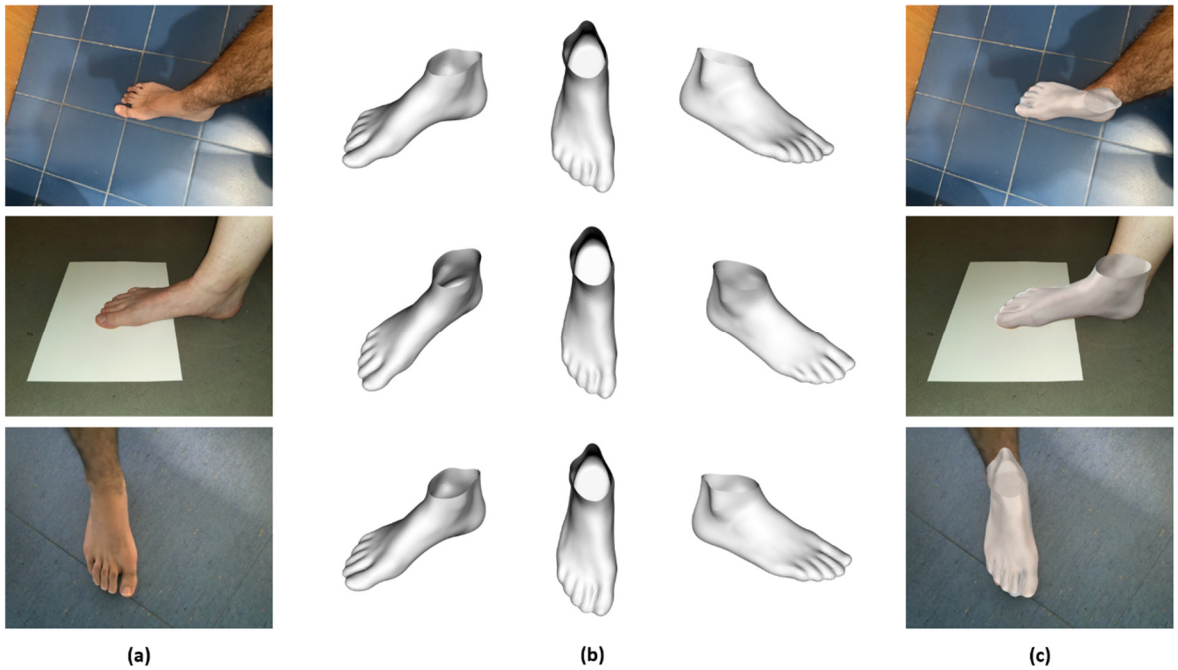

Figure 5: Examples of reconstructions: (a) input image, (b) 3D reconstructed foot and (c) 3D foot projected on the image

6.4. Real-Time Inference

The model's ability to perform real-time inference was also evaluated in Table 2. During testing, the model consistently produced accurate 3D reconstructions from 2D images within milliseconds, demonstrating its potential for use in dynamic applications where quick turnaround is essential.

Device	Model type	Inference Time
Computer CPU (11th Gen Intel(R) Core(TM) i7-11800H)	Onnx Model	29 ms
Computer GPU (NVIDIA GeForce RTX 3070)	Pytorch Model	25 ms
Smartphone GPU (Mali-G72 MP18)	TF Lite Model	40 ms

Table 2: Inference time

7. Conclusions

In this study, we introduced a novel neural network-based approach for accurate 3D foot reconstruction from a single 2D image. Our method leverages the power of deep learning combined with statistical shape modeling through Principal Component Analysis (PCA) to infer 3D foot morphology. The integration of binary segmentation masks and 2D keypoint estimation further enhances the precision of the model, ensuring detailed and anatomically accurate reconstructions.

The training process, utilizing both synthetic and augmented real-world datasets, has demonstrated the model's robustness and adaptability across diverse scenarios. The validation on a large dataset resulted in a Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) of 0.9 mm, underscoring the method's high accuracy.

The real-time capabilities and high precision achieved with a single image open up new possibilities for augmented reality applications, such as virtual try-on solutions for footwear and innovative 3D foot capture experiences. The ability to provide a complete foot shape rather than just a bounding box, along with precise position and rotation information, enables the development of more accurate, fluid, and realistic virtual try-on applications compared to current solutions.

Moreover, advancements in foot capture using mobile devices are promising. Our method enables a faster, smoother, and more natural capture process compared to existing systems that rely on multiple photographs. This could significantly enhance accuracy by incorporating additional images or selecting optimal samples during the capture process. Additionally, real-time reconstruction facilitates synchronization between model generation and smartphone accelerometers. This allows for metric scale calculation without the need for known-size reference objects, thereby increasing the methodology's versatility. This represents a significant improvement for both clinical applications and consumer product selection or customization, including insoles, footwear, and orthotics.

In conclusion, our proposed approach marks a significant advancement in 3D foot reconstruction from 2D images, combining efficiency, accuracy, and versatility. Future work may explore the extension of this method to other body parts and further optimization for specific applications, enhancing its utility and impact in the field of computer vision and beyond. For example, using a dataset captured with Move4D system [40], we are extending this methodology to articulated full-body models, which could have even greater implications for augmented reality applications beyond virtual try-ons and 3D human body capture. Potential applications include advancements in healthcare and sports, demonstrating the broad impact and versatility of our approach.

References

[1] N. Cho, S. Kim, D.-J. Kwon, and H. Kim, “The prevalence of hallux valgus and its association with foot pain and function in a rural Korean community,” *J. Bone Joint Surg. Br.*, vol. 91, pp. 494–8, Apr. 2009, doi: 10.1302/0301-620X.91B4.21925.

[2] L. Ceyssens, R. Vanelderen, P. Malliaras, and B. Dingenen, “Biomechanical Risk Factors Associated with Running-Related Injuries: A Systematic Review,” *Sports Med.*, vol. 49, Jul. 2019, doi: 10.1007/s40279-019-01110-z.

[3] B. M. Nigg, M. A. Nurse, and D. J. Stefanyshyn, “Shoe inserts and orthotics for sport and physical activities.,” *Med. Sci. Sports Exerc.*, vol. 31 7 Suppl, pp. S421-8, 1999.

[4] E. Parrilla *et al.*, “Low-cost 3D foot scanner using a mobile app,” *Footwear Sci.*, vol. 7, no. sup1, pp. S26–S28, Jun. 2015, doi: 10.1080/19424280.2015.1038308.

[5] A. Ballester *et al.*, “Fast, Portable and Low-Cost 3D Foot Digitizers: Validity and Reliability of Measurements,” presented at the Proc. of 3DBODY.TECH 2017 - 8th Int. Conference and Exhibition on 3D Body Scanning and Processing Technologies, Montreal QC, Canada, 11-12 Oct. 2017, Oct. 2017. doi: 10.15221/17.218.

[6] A. Ferrari *et al.*, “Quantitative comparison of five current protocols in gait analysis.,” *Gait Posture*, vol. 28 2, pp. 207–16, 2008.

[7] S. Telfer, K. Gibson, K. Hennessy, M. Steultjens, and J. Woodburn, “Computer-Aided Design of Customized Foot Orthoses: Reproducibility and Effect of Method Used to Obtain Foot Shape,” *Arch. Phys. Med. Rehabil.*, vol. 93, pp. 863–70, May 2012, doi: 10.1016/j.apmr.2011.12.019.

[8] “Scan&Fit – Base Protection World.” Accessed: Jul. 26, 2024. [Online]. Available: https://www.baseprotection.com/scanfit/

[9] “Fit | Dodge Ski Boots.” Accessed: Jul. 26, 2024. [Online]. Available: https://dodgeskiboots.com/fit/

[10] “Home,” Feetz SizeMe Shoes. Accessed: Jul. 26, 2024. [Online]. Available: http://www.feetz.co/

[11] “helu one^2 carbon,” Hezo Cycling. Accessed: Jul. 26, 2024. [Online]. Available: https://www.hezo-cycling.com/products/helu-one

[12] “Shooax.” Accessed: Jul. 26, 2024. [Online]. Available: https://shooax.cz/#production

[13] Z. Chai *et al.*, “HiFace: High-Fidelity 3D Face Reconstruction by Learning Static and Dynamic Details,” Aug. 23, 2023, *arXiv*: arXiv:2303.11225. doi: 10.48550/arXiv.2303.11225.

[14] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, “One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer,” Mar. 28, 2023, *arXiv*: arXiv:2303.16160. doi: 10.48550/arXiv.2303.16160.

[15] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, and R. Taniguchi, “TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell.” 2020.

[16] K. Lin, L. Wang, Y. Jin, Z. Liu, and M.-T. Sun, “Learning Nonparametric Human Mesh Reconstruction from a Single Image without Ground Truth Meshes.” 2020.

[17] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “DeepHuman: 3D Human Reconstruction from a Single Image.” 2019.

[18] T. Zhang, B. Huang, and Y. Wang, “Object-Occluded Human Shape and Pose Estimation From a Single Color Image,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[19] H. Xu, E. G. Bazavan, A. Zanfir, B. Freeman, R. Sukthankar, and C. Sminchisescu, “GHUM and GHUML: Generative 3D Human Shape and Articulated Pose Models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*, 2020, pp. 6184–6193. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Xu_GHUM__GHUML_Generative_3D_Human_Shape_and_Articulated_Pose_CVPR_2020_paper.html

[20] A. Maćkiewicz and W. Ratajczak, “Principal components analysis (PCA),” *Comput. Geosci.*, vol. 19, no. 3, pp. 303–342, Mar. 1993, doi: 10.1016/0098-3004(93)90090-R.

[21] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image.” 2016.

[22] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end Recovery of Human Shape and Pose.” 2018.

[23] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, “Learning to Estimate 3D Human Pose and Shape from a Single Color Image.” 2018.

[24] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, “SCAPE: Shape completion and animation of people,” *ACM Trans. Graph. TOG*, vol. 24, pp. 408–416, Jul. 2005, doi: 10.1145/1073204.1073207.

[25]H. Joo, T. Simon, and Y. Sheikh, “Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies.” 2018.

[26] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross, “Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks,” Jul. 2017, pp. 5504–5514. doi: 10.1109/CVPR.2017.584.

[27] J. Sun, M. Ovsjanikov, and L. Guibas, “A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion,” *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009, doi: 10.1111/j.1467-8659.2009.01515.x.

[28] Y. You, H. Liu, T. Wang, W. Li, R. Ding, and X. Li, “Co-Evolution of Pose and Mesh for 3D Human Body Estimation from Video,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 14963–14973.

[29] 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,” Aug. 03, 2020, *arXiv*: arXiv:2003.08934. doi: 10.48550/arXiv.2003.08934.

[30] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy Networks: Learning 3D Reconstruction in Function Space,” Apr. 30, 2019, *arXiv*: arXiv:1812.03828. doi: 10.48550/arXiv.1812.03828.

[31] L. Yariv *et al.*, “Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance,” Oct. 25, 2020, *arXiv*: arXiv:2003.09852. doi: 10.48550/arXiv.2003.09852.

[32] Y. Tian, H. Zhang, Y. Liu, and L. Wang, “Recovering 3D Human Mesh from Monocular Images: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15406–15425, Dec. 2023, doi: 10.1109/TPAMI.2023.3298850.

[33] J. Li *et al.*, “Human Pose Regression with Residual Log-likelihood Estimation,” Jul. 31, 2021, *arXiv*: arXiv:2107.11291. doi: 10.48550/arXiv.2107.11291.

[34] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, “Image Data Augmentation for Deep Learning: A Survey,” Nov. 05, 2023, *arXiv*: arXiv:2204.08610. doi: 10.48550/arXiv.2204.08610.

[35] G. Varol *et al.*, “Learning from Synthetic Humans,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 4627–4635. doi: 10.1109/CVPR.2017.492.

[36] Z. Yang *et al.*, “SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling,” Sep. 11, 2023, *arXiv*: arXiv:2303.17368. doi: 10.48550/arXiv.2303.17368.

[37] A. Sengupta, I. Budvytis, and R. Cipolla, “Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild,” Sep. 22, 2020, *arXiv*: arXiv:2009.10013. doi: 10.48550/arXiv.2009.10013.

[38] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Jan. 29, 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

[39] K. H. Zou *et al.*, “Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index,” *Acad. Radiol.*, vol. 11, no. 2, pp. 178–189, Feb. 2004, doi: 10.1016/S1076-6332(03)00671-8.

[40] E. Parrilla *et al.*, *MOVE 4D: Accurate High-Speed 3D Body Models in Motion*. 2019, p. 32. doi: 10.15221/19.030.