FIND: An Unsupervised Implicit 3D Model of Articulated Human Feet

Oliver Boyne, James Charles, Roberto Cipolla
University of Cambridge

BMVC 2022

UNIVERSITY OF CAMBRIDGE

Motivation

- Modelling feet is useful for shoe fitting and orthotics
- Accurate generative models of bodies [1], hands [2] and faces [3] have been well developed
- Foot models are a relatively unexplored category – typical shape reconstruction uses point clouds [4] or low-resolution PCA models [5]
- Producing foot models is challenging due to limited available data

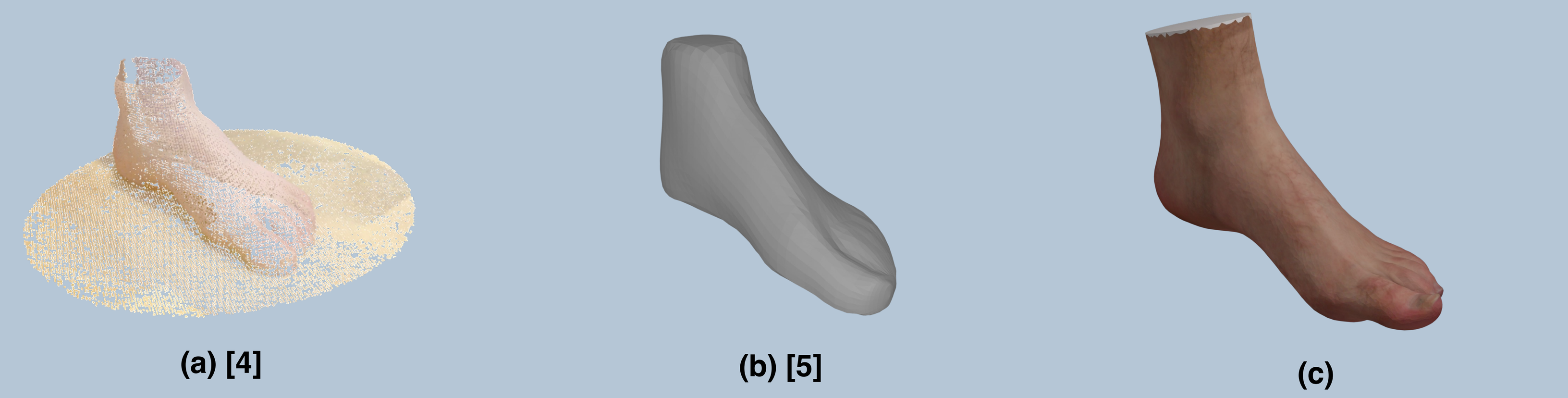

Figure 1: (a) Point cloud reconstruction and (b) a PCA model are unable to capture the geometry and texture of (c) a high resolution foot scan

Contributions

Figure 2: Outputs of the FIND model

Figure 3: Foot3D scans with pose descriptions

- **FIND** (Foot Implicit Neural Deformation) model which generates **explicit**, **textured** feet with **pose, shape and texture**
  - Unsupervised shape/pose disentanglement
  - Unsupervised part-based learning
- **Foot3D** dataset of high resolution, textured foot scans in a variety of poses

Method - FIND Model

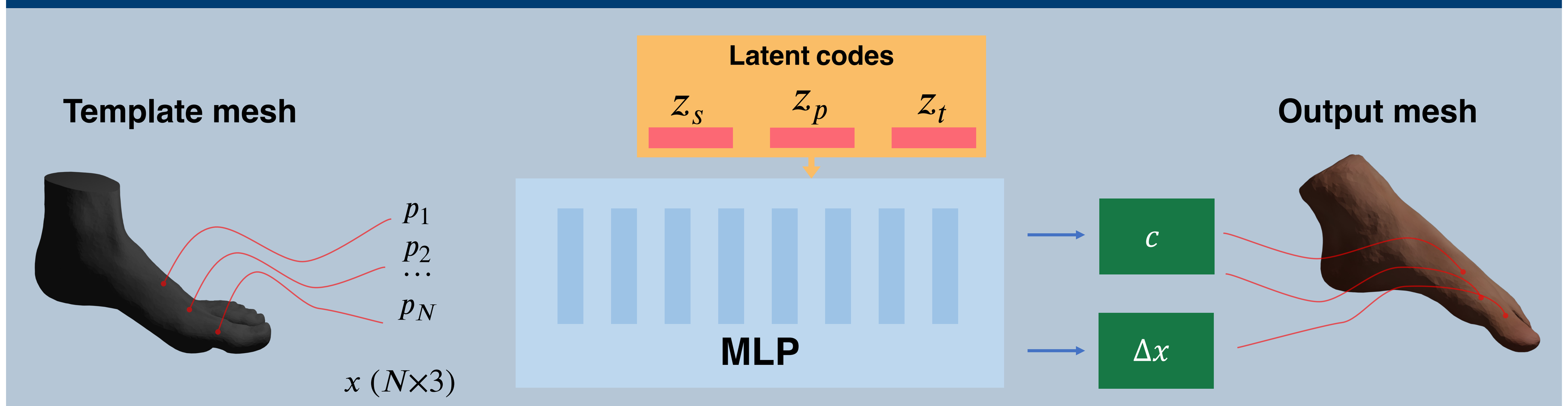

Figure 4: FIND model overview

- Given latent codes z_s (shape), z_p (pose), z_t (texture)
- Sample points x on the surface of template mesh
- Feed positional encoding $\gamma(x)$ through MLP F to predict colour c and displacement Δx

$$F\big(\gamma(x), z_s, z_p, z_t\big) \rightarrow (\Delta x, c)$$

- Unsupervised pose representation learning
  - Constraint: feet of same identity have same z_s
  - Contrastive loss: similar poses have similar z_p; different poses have different z_p

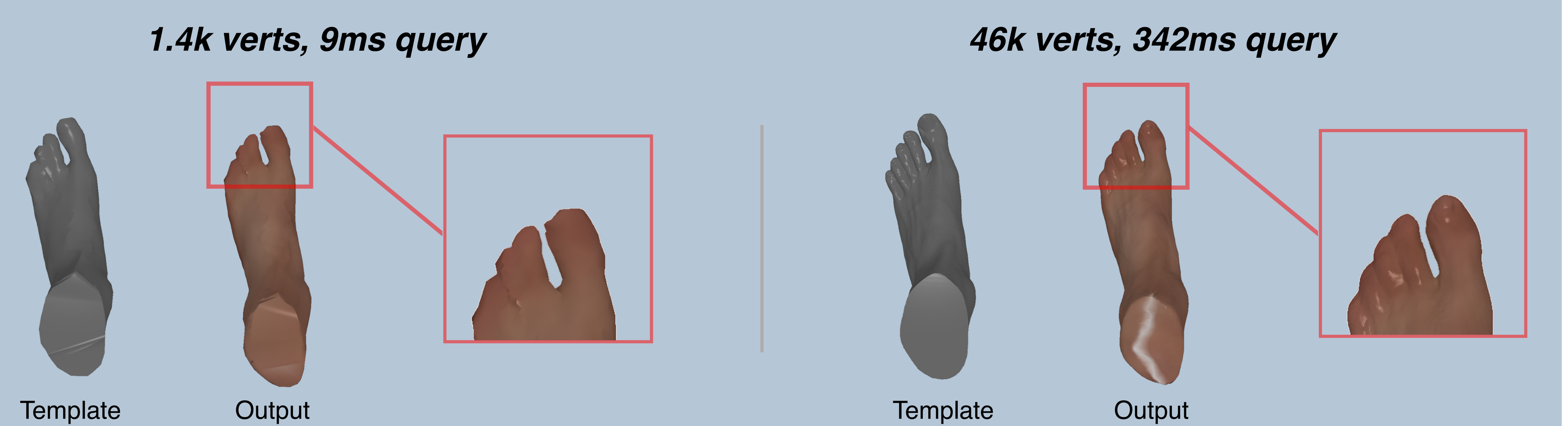

Figure 5: Multi-resolution capability of the model. For speed or memory critical applications, a low resolution template mesh can be used.

Method - Learning parts

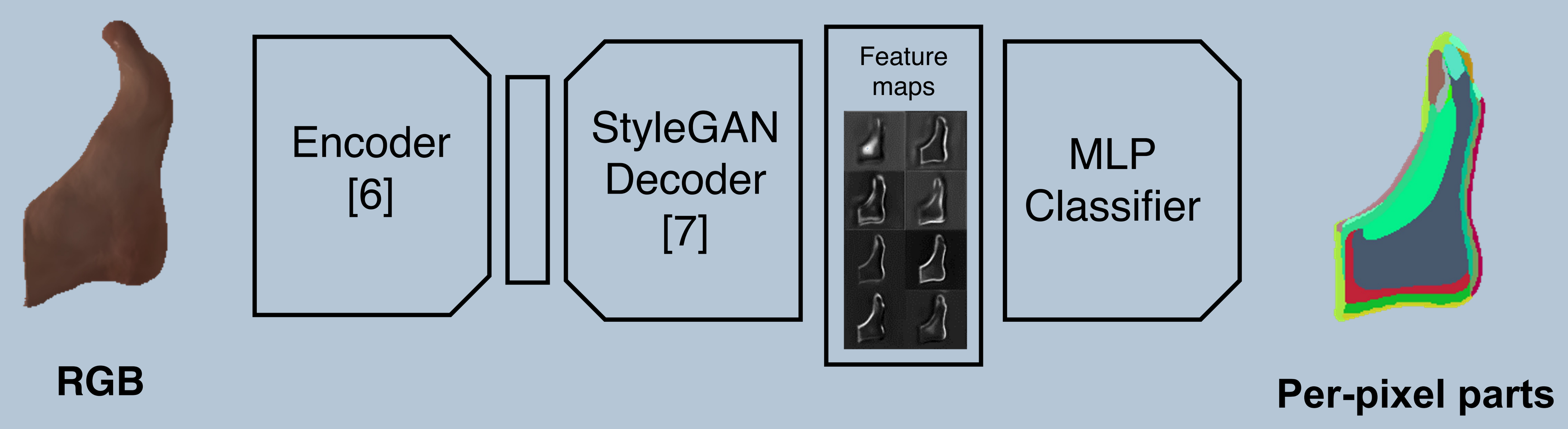

Figure 6: Pipeline for predicting per-pixel parts from an input image

- StyleGAN [7] generates synthetic foot images
- Encode [6] foot images to StyleGAN style codes
- k-means clustering on StyleGAN feature maps produces 'part' segmentations
- Train MLP classifier to predict these parts
- Fully differentiable image-to-parts pipeline (Figure 6)
- At train time, use pipeline to learn parts directly on template mesh of FIND
- For inference on 2D images, use cross entropy between image-to-parts pipeline and projected 3D FIND parts

Experimental Results

- **3D evaluation**: model evaluated by fitting to Foot3D validation scans, with 3D chamfer loss

Model	Trained on	Chamfer, μm ↓	Keypoint, mm ↓	IoU ↑
SUPR [8]	4D foot scans	48.0	11.2	0.756
PCA [9]	Foot3D	11.2	15.7	0.892
FIND	Foot3D	**7.3**	**5.9**	**0.931**

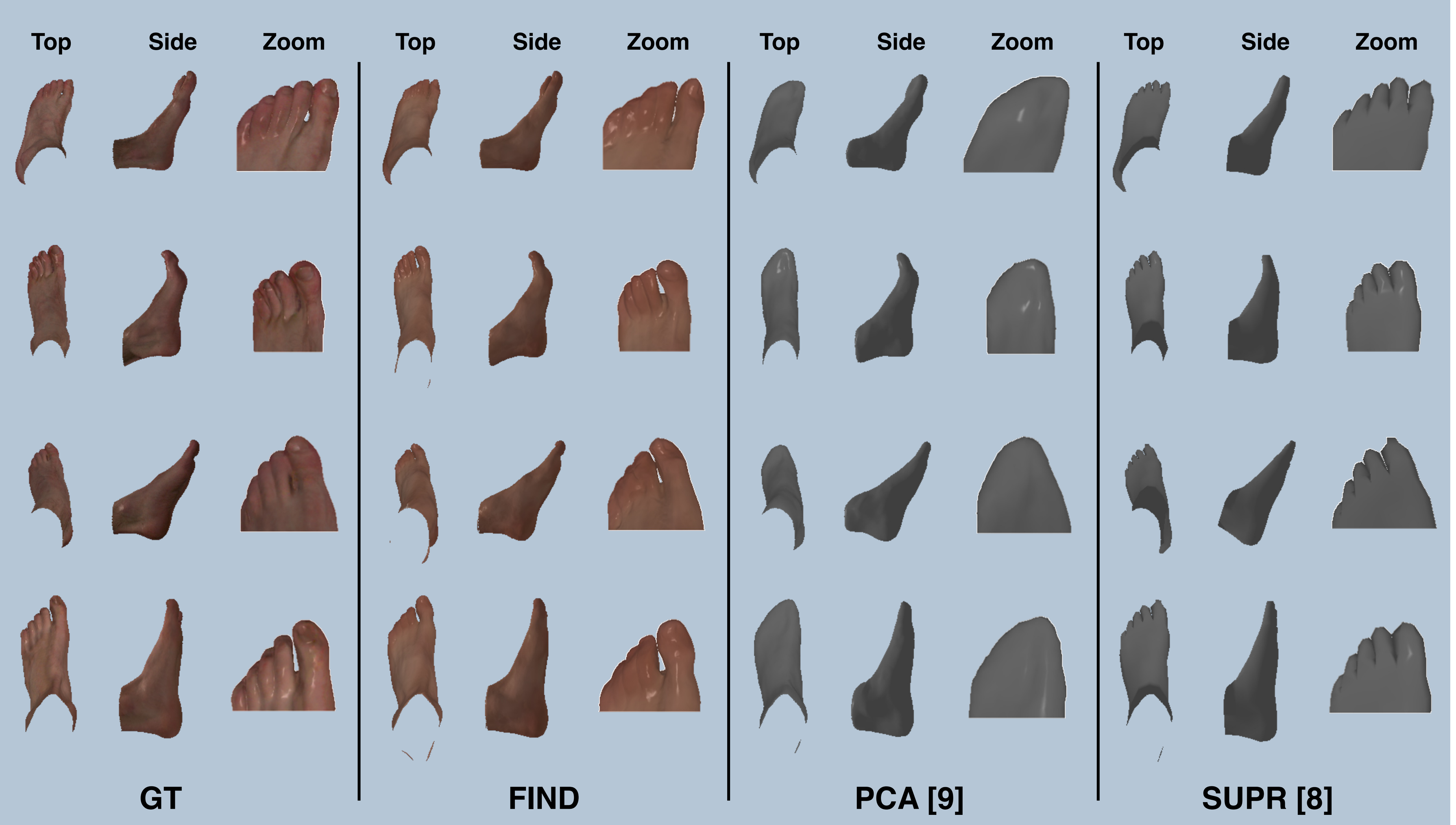

Figure 7: Qualitative results of 3D fitting to validation scans

- **2D evaluation**: model fitted to synthetic renders of Foot3D scans, using (i) silhouette loss only, (ii) silhouette + VGG [10] perceptual loss, and (iii) silhouette + cross-entropy loss using our learned foot parts

	2 view		5 view	
Optimisation loss	Chamfer, μm ↓	Keypoint, mm ↓	Chamfer, μm ↓	Keypoint, mm ↓
Sil	81.8	14.4	16.8	7.7
Sil + VGG [10]	78.7	13.1	15.9	7.3
Sil + CE Loss	**45.8**	**10.3**	**15.7**	**6.4**

References

[1] Loper et al. SMPL: A Skinned Multi-Person Linear Model. ACM TOG 2015
[2] Li et al. Learning a model of facial shape and expression from 4D scans. ACM TOG 2017
[3] Romero et al. Embodied hands: Modeling and capturing hands and bodies together. ACM TOG 2017
[4] https://www.xesto.io
[5] https://www.snapfeet.io
[6] Alaluf et al., ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. ICCV 2021
[7] Richardson et al., Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. CVPR 2021
[8] Osman et al. SUPR: A Sparse Unified Part-Based Human Representation. ECCV 2022
[9] Yang et al. FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation. CVPR 2018
[10] Simonyan et al. Very deep convolutional networks for large-scale image recognition. ICLR 2015

ollieboyne.github.io/FIND

Dataset ● Code ● Web demo

Email: ob312@cam.ac.uk