FOCUS - Multi-View Foot Reconstruction From Synthetically Trained Dense Correspondences

Oliver Boyne Roberto Cipolla
Department of Engineering, University of Cambridge, U.K.
{ob312, rc10001}@cam.ac.uk

Abstract

Surface reconstruction from multiple, calibrated images is a challenging task - often requiring a large number of collected images with significant overlap. We look at the specific case of human foot reconstruction. As with previous successful foot reconstruction work, we seek to extract rich per-pixel geometry cues from multi-view RGB images, and fuse these into a final 3D object. Our method, FOCUS, tackles this problem with 3 main contributions: (i) SynFoot2, an extension of an existing synthetic foot dataset to include a new data type: dense correspondence with the parameterized foot model FIND; (ii) an uncertainty-aware dense correspondence predictor trained on our synthetic dataset; (iii) two methods for reconstructing a 3D surface from dense correspondence predictions: one inspired by Structure-from-Motion, and one optimization-based using the FIND model. We show that our reconstruction achieves state-of-the-art reconstruction quality in a few-view setting, performing comparably to state-of-the-art when many views are available, runs substantially faster, and can run without a GPU. We release our synthetic dataset to the research community. Code is available at: https://github.com/OllieBoyne/FOCUS

1. Introduction

Accurate 3D reconstruction of human body parts from images is a challenging computer vision task, of significant interest to the health, fashion and fitness industry. In this paper, we address the problem of reconstructing a human foot accurately from multiple views. Shoe retail, orthotics and health monitoring can all be improved with accurate models of the foot, and the growing digital markets for these applications has created a demand for recovering such models from mobile phone images captured by ordinary users.

Existing approaches to foot reconstruction include: (i) specialized scanning equipment [1, 2]; (ii) reconstruction of unoriented point clouds from depth maps or phone-based sensors [4, 26]; (iii) photogrammetry pipelines, such as COLMAP, of Structure-from-Motion (SfM) followed by Multi-View Stereo (MVS) [32, 33]; and (iv) fitting generative models to silhouettes, keypoints and surface normals [13, 22].

These methods come with substantial drawbacks: (i) expensive scanning equipment is not accessible to most consumers; (ii) phone-based sensors are limited in availability and ease of use, and noisy, unoriented point clouds are difficult to use for desired applications such as rendering and

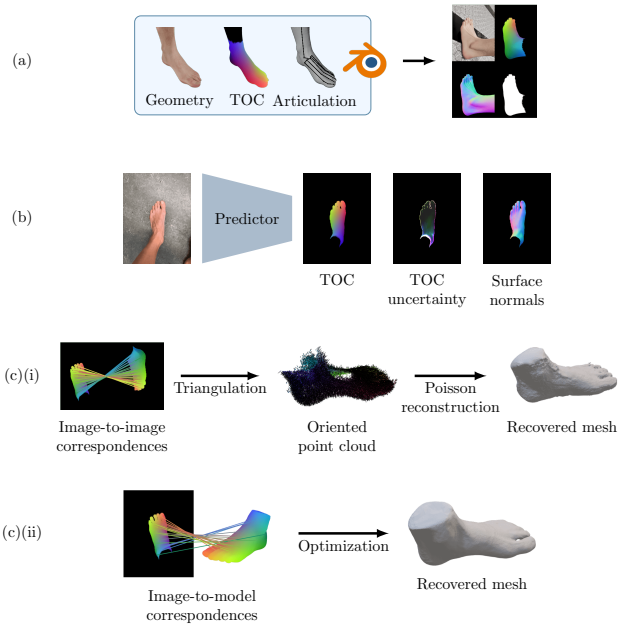

Figure 1. **Method overview.** (a) We use Blender [15] to render articulated high resolution meshes, with dense correspondences (TOC) to the generative FIND [12] model. (b) We train a model to predict TOCs and surface normals on real images. (c) We combine these predictions together in a multi-view setting via two methods to yield accurate surface reconstructions: (i) FOCUS-SfM, a Structure-from-Motion based approach; and (ii) FOCUS-O, a model fitting, optimization-based approach.

taking measurements; (iii) SfM and MVS require a large number of input views and favourable lighting conditions; and (iv) current optimization-based approaches use silhouettes, keypoints and surface normals, which only provide a limited representation of the full set of available cues in the image that can be used for accurate surface reconstruction.

To this end, we introduce FOCUS, **F**oot **O**ptimization via **C**orrespondences **U**sing **S**ynthetics - which improves on current multi-view reconstruction pipelines by predicting per-pixel correspondences relative to the FIND [12] parameterized foot model. We find this representation provides a much stronger geometric grounding than solely silhouettes [22] and surface normals [13] seen in prior foot reconstruction work, allowing for a more direct training signal from which to reconstruct a foot model. We outline our method in Figure 1, and our key contributions are as follows:

- In order to learn per-pixel correspondences, we extend prior research into foot synthetic data to release SynFoot2, a **large scale synthetic dataset** of 100,000 photorealistic foot images, coupled with accurate surface normals and dense correspondence labels. We drastically increase the background and lighting variety compared to *SynFoot*, and introduce articulation into the scans to provide a greater pose variation.
- We introduce **Template Object Coordinates** (TOCs), normalized coordinates in the space of the FIND parameterized foot model, as a new representation of foot geometry. We train a predictor to jointly predict these per-pixel correspondences, in addition to a corresponding uncertainty. We follow a similar approach to prior work, to successfully train the network solely on synthetic data, and show plausible predictions on in-the-wild images.
- We introduce FOCUS-SfM, a method for fusing multiple views of correspondence predictions into a single oriented point cloud, by matching correspondences and triangulating, similar to Structure-from-Motion. We use Poisson reconstruction [20] to convert the oriented point cloud to a final mesh. This method does not require a GPU, requires fewer views than COLMAP, and is substantially faster than previous work.
- We also introduce FOCUS-O, a method for directly optimizing the parameterized FIND [12] model to these dense correspondence images, producing a watertight, parameterized mesh as output. This method directly uses the predicted correspondence uncertainty, and is capable of accurate reconstruction on as few as 3 views.

2. Related Work

Synthetic dataset generation. Synthetic rendering has become a growing source for data generation in the computer vision community. As photorealistic rendering capabilities have improved, synthetic pipelines have been used to produce high quality, large scale datasets, that are less expensive to scale than manually labelled datasets, and can often be more accurate. These pipelines are also useful for tasks that are difficult or impossible for human labellers, such as per-pixel labelling. Existing research has shown the viability of mostly, or entirely, synthetic data in training for complex downstream tasks - examples for human body reconstruction include bodies [37], faces [8, 40], eyes [39], and feet [13].

Multi-view reconstruction. In order to recover geometry from multiple images, the relative camera positions and camera internal parameters must be known. These can be obtained directly from measurements of the image capturing device using onboard Inertial Measurement Units (IMUs), or can be recovered via sparse 3D reconstruction from Structure From Motion (SfM) [32].

To recover the surface geometry, a common method is Multi-View Stereo (MVS) [33], a process of optimizing depth and normal maps across views, from which an oriented point cloud can be recovered. From this, a mesh is constructed using a surface reconstruction algorithm [20]. COLMAP [32, 33] is a popular implementation of this full pipeline.

In recent years, neural rendering for surface reconstruction has grown in interest [27, 41]. In these methods, a neural representation of a 3D scene is trained, which, when rendered through some differentiable process, accurately reconstructs reference views. These methods often require large amounts of training time and compute, and large numbers of input views for accurate reconstructions.

Dense correspondences. For pose and shape reconstruction, sparse correspondences, or keypoints, are often used for registration. Obtaining accurate keypoint labels is fast and easy to explain to human labellers.

To fully reconstruct an accurate surface, dense correspondences can be incredibly useful - a mapping from every pixel in one image to a pixel in another. Some implementations treat this as a generic, image pair matching problem - often called *optical flow*. Finding this flow field for image pairs has been shown to be solvable via both optimization [24, 35] using image features, or by training a model to directly predict this flow [31, 34].

Other implementations seek to match all pixels of an object in an image to that object's local space, normalized to a unit cube. This approach, introduced by Wang et al. [38] as Normalized Object Coordinates (NOCs), has been used to reconstruct object size and pose from single images, and by Gümeli et al. [17] for multi-view joint object pose and camera registration.

For matching an object of a known category, it can be more advantageous to be able to map any number of images to some canonical object space - this allows for more

effective fusion of multiple views, and for matching across a wider range of viewpoints. Obtaining the data necessary to learn this canonical mapping can be difficult. For human body estimation, Güler et al. [16] inferred dense correspondences from sparse human labelling and part matching. Taylor et al. [36] directly learn dense correspondences from depth images via regression forests.

Zeng et al. [42] learn to predict a mapping between pixels of an image of a human, and a UV map of a human body mesh, and task a regressor to directly recover a 3D surface from this prediction on a single image.

We also learn a mapping directly to a parameterized mesh. We do this entirely synthetically, and are able to leverage these correspondences in a multi-view setting to obtain an accurate reconstruction.

Human foot reconstruction. Obtaining accurate models of the human foot is of significant interest to the footwear and orthotics industries. The prevalence of mobile phones and digital shopping and health has increased the demand for these reconstruction methods to be made possible with data collected from consumer devices.

Some methods collect point cloud data [4, 26], which can be achieved with LiDAR or structured light sensors available on certain mobile phones. Such sensors are not universally available, and often these point clouds are noisy and not capable of providing a detailed, realistic foot surface.

Other methods instead seek to fit parameterized models of feet to input data. Earlier works built Principal Component Analysis (PCA) models of feet from sampling vertices from photogrammetry [22] and scanners [6]. Kok et al. [22] also fit their PCA model to predicted image silhouettes.

The PCA approach has been improved in recent years by Osman et al. [29], with SUPR, a PCA model of the human foot to be combined with the SMPL [25] full body model for the task of expressive, full body reconstruction.

Another approach is the FIND model [12], a non-linear generative model of the foot. Rather than PCA, FIND uses an implicit network to deform a template mesh per-vertex to a target pose and shape. With this model, Foot3D was released - a collection of high resolution foot scans.

Recently, FOUND [13] introduced a method of optimizing the parameters of the FIND model to match with predicted surface normals in 2D in a multi-view setting. Due to a lack of available real surface normal data, the surface normal predictor was trained in a synthetic setting, generalizing well to in-the-wild images.

3. Method

3.1. Template Object Coordinates

We seek to define some dense correspondence over a foot surface. Many dense correspondence predictors, [16, 36, 42], especially for human body prediction, use a 2D parameterization of the surface, such as a UV mapping. We choose to use a 3D parameterization, for three reasons: (i) ease of construction; (ii) a more intuitive uncertainty representation; and (iii) consistency with the coordinate representation used in the FIND [12] model, which is key to both our synthetic data and reconstruction approaches.

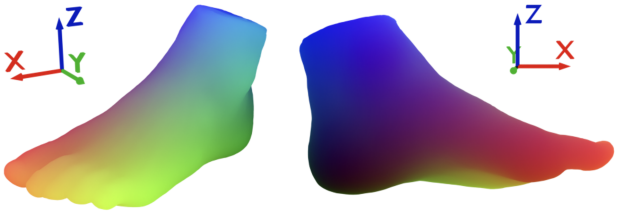

Figure 2. **TOC definition.** Template Object Coordinates (TOCs), shown on the template of the FIND mesh. RGB values correspond to XYZ, normalized to 0-1 within the template space.

(a) (b) (c)

Figure 3. **Models for rendering.** One of 8 foot models used for the synthetic dataset. The mesh has (a) geometry and texture, (b) a TOC mapping to the FIND template model, and (c) a skeleton used for articulation.

We therefore define **Template Object Coordinates** (TOCs), denoted as $\mathbf{t}$. Visualized in Figure 2, TOC values represent points in the 3D local space of the FIND model's template mesh, normalized to the template's bounding box.

Where Normalized Object Coordinates (NOCs) simply map to points in an object's local space, TOCs map points from any deformed, articulated foot directly to the template's local space. This is essential to directly fit the FIND model to TOC predictions, and allows for our per-pixel predictive network in Section 3.3 to predict correspondences agnostic to pose, shape and identity.

3.2. SynFoot2

We extend SynFoot [13], a large scale synthetic dataset for feet: we add articulated feet, our new TOC representation, and increased background and lighting diversity.

Articulation. We manually add a skeleton to all 8 meshes in the original dataset, in line with anatomical diagrams [19]. We use Blender's automatic weight calculation [9] to handle the vertex weighting.

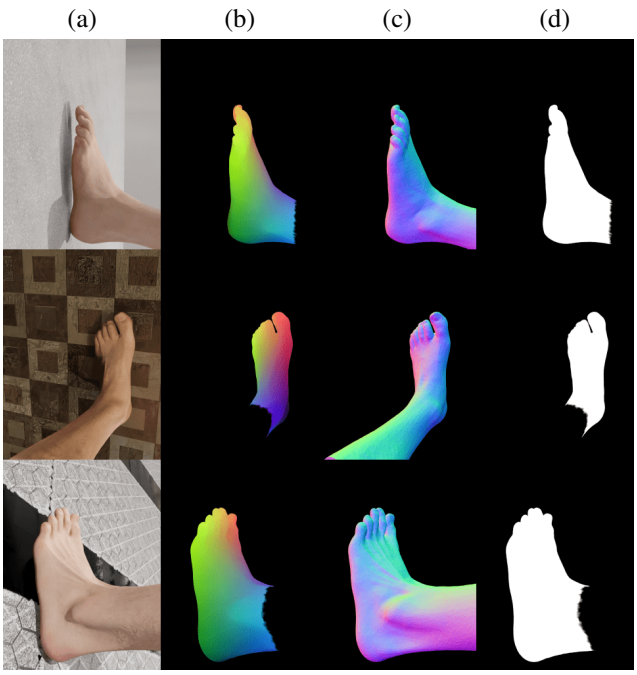

Figure 4. **SynFoot2 examples.** We show (a) RGB, (b) TOC, (c) surface normals, and (d) segmentation masks. Further examples are included in the supplementary material.

We identify 5 methods of articulation described in foot literature [19]: dorsiflexion/plantarflexion (pitch), lateral/medial rotation (yaw), inversion/eversion (roll), toe extension/flexion (up/down), and toe abduction/adduction (outwards/inwards). We randomly sample these poses and apply combinations of them to our meshes to provide variation in articulation in the synthetic data.

TOCs. We render TOCs for all meshes, by first fitting the FIND mesh to each model as in the original FIND paper [12], and using this to find a vertex-to-vertex mapping between each mesh and the FIND template space. This provides us with per-vertex TOC values for each mesh to be used for our synthetic data, as in Figure 3.

Rendering. We use the BlenderSynth [13] package to render the dataset. We drastically increase the quantity and variety of HDRIs and background textures compared to SynFoot [13], increasing the number of HDRIs from 14 to 733, and background textures from 34 to 541 using Poly Haven assets [3]. We add a new shader to render TOCs. We render 100,000 images at 480 x 640 resolution. Figure 4 shows some examples of our synthetic dataset.

3.3. Training a predictor

We start with a model used to make coarse surface normal and uncertainty predictions, and uncertainty-guided pixel-wise refinements, as in [5, 7]. We modify the output prediction heads to tackle our downstream tasks.

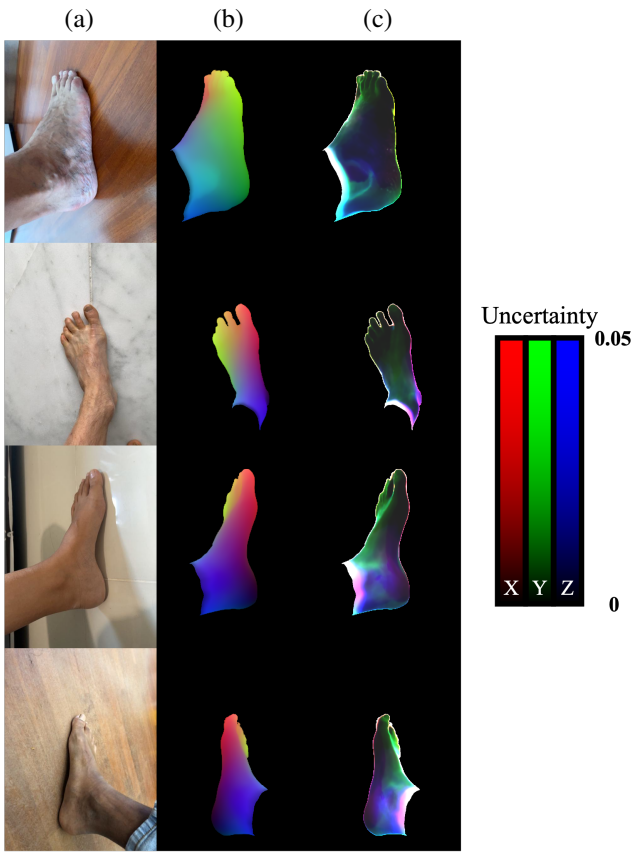

Figure 5. **TOC in-the-wild predictions.** Predictions on real images, showing (a) RGB input, (b) TOC $\mathbf{t}$, (c) TOC uncertainty $\boldsymbol{\sigma}_\mathbf{t}$. Further examples are included in the supplementary material.

First, we add a head to predict a binary segmentation heatmap for the foot using Binary Cross-Entropy loss. At inference time, we threshold all predictions according to this heatmap being larger than 0.5.

Next, we add heads to predict a probability distribution for the TOC values. We model our TOC prediction as a normal distribution in XYZ,

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{t}, \boldsymbol{\sigma}_\mathbf{t}^2), \tag{1}$$

and have one head to predict $\boldsymbol{\mu}_\mathbf{t}$, and another to predict $\log \boldsymbol{\sigma}_\mathbf{t}^2$, independently in each axis XYZ. We combine these predictions in an uncertainty-aware loss during training of the predictor, which minimizes the negative log-likelihood of the TOC predictions,

$$\mathcal{L}_{\text{TOC}} = \left\| \frac{(\boldsymbol{\mu}_\mathbf{t} - \mathbf{t}_{\text{gt}})^2}{\boldsymbol{\sigma}_\mathbf{t}^2} \right\|_2 + \log \boldsymbol{\sigma}_\mathbf{t}^2. \tag{2}$$

We show examples of TOC inference on in-the-wild images in Figure 5.

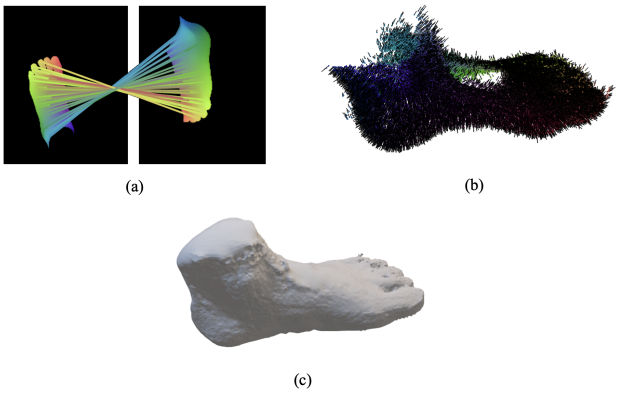

Figure 6. **FOCUS-SfM overview.** (a) We find correspondences between images by matching TOC values; (b) we triangulate and collect normals across views to construct an oriented point cloud; (c) we use Poisson surface reconstruction to form a final mesh.

3.4. FOCUS-SfM

Given these TOC predictions, we now wish to reconstruct a surface. The first of our two proposed methods for reconstruction, FOCUS-SfM, is outlined in Figure 6 and takes inspiration from traditional Structure-from-Motion.

We predict TOCs, segmentation masks and surface normals on N images of a captured foot with known camera extrinsics and intrinsics. FOCUS-SfM aims to identify correspondences between these predictions, and use triangulation to reconstruct an oriented point cloud.

Sampling. For each image, we sample P points within the mask of prediction. For each point, we collect the value of the pixel position, TOC, and surface normal.

Pixel-level correspondences. We now want to find correspondences between images, where a 'correspondence' refers to an exact TOC value. Each correspondence might only appear in a subset of images. We employ an efficient approach using nearest neighbor sampling. For each image, we structure all TOC values as a KD-tree [11] and, for each set of P correspondences, look up the nearest neighbor in each image by $\ell 2$ distance.

Subpixel correspondences. We achieve more accurate correspondences by matching at the sub-pixel level. We use bilinear interpolation to upscale the 3x3 patch around each found correspondence by a factor of 8, and then search within this window for a better match. This search is implemented in Cython [10] for performance.

Once we have the match, we capture the surface normal at each point. We only consider a correspondence found if the TOC value is within 0.002 $\ell 2$ distance of the original sampled value.

The result of this is $C \leq NP$ correspondences, each containing a subset of the N views, and pixel positions for each view.

Triangulation. For each correspondence, we triangulate across all views in which that correspondence is present, using the known camera parameters, and the Direct Linear Transformation [18]. This provides a point cloud of triangulated correspondences.

Filtering. We filter the collected point cloud via three methods: (i) removing points whose average reprojection error is above a threshold; (ii) removing all points below the floor (Z=0); (iii) statistical outlier removal [23].

Normal aggregation. We also collect normals along the correspondences. To provide a normal estimate, we convert all normals to spherical coordinates $(1, \theta, \phi)$, and average over θ and ϕ to reach a consensus normal.

Poisson surface reconstruction. Now that we have an oriented point cloud, we use Screened Poisson Reconstruction [20] in Meshlab [14] to reconstruct a surface.

Implementation details. The hyperparameters chosen in this process, such as for filtering and Poisson reconstruction, can be found in the supplementary material.

3.5. FOCUS-O

The second approach we introduce, FOCUS-O, takes inspiration from FOUND [13] - fitting a parameterized model directly to predictions made in image space. We seek to optimize the global transformation (r, s, t) and FIND shape and pose embeddings (z_s, z_p).

Rather than use differentiable rendering, as in FOUND, the TOC representation allows us to sample points on the image, and optimize the FIND model such that the corresponding point on the FIND model projects onto the same pixel position. This is essentially a keypoint loss, where any number of keypoints can be sampled at arbitrary locations in image space.

Sampling. As in Section 3.4, we sample P $\mathbf{t}$ values for each of the N images within the mask of prediction, as well as recording their pixel position $\boldsymbol{p}$.

The TOC values $\mathbf{t}$ are in normalized space - we convert them to FIND space, $\mathbf{t}'$, by mapping to the axis aligned bounding box of the FIND template mesh.

FIND model. As defined in [12], the FIND model maps a 3D point on the surface of a template mesh $\mathbf{x_1}$ to a deformed mesh, under shape and pose embeddings (z_s, z_p), and global transformation (r, s, t), to a deformed point $\mathbf{x_2}$,

$$\mathbf{x_2} = F(\mathbf{x_1}, z_s, z_p, r, s, t) \tag{3}$$

Projection. Under our FIND and camera models, we want to project our $\mathbf{t}'$ estimates onto the image plane. Given a function that maps a world point to pixel space under a given camera model, f, and the FIND model F, a reprojected point in 2D space can be calculated,

$$\hat{\boldsymbol{p}} = f\left(F\left(\mathbf{t}', z_s, z_p, r, s, t\right)\right). \tag{4}$$

For our experiments, we use the default camera model of COLMAP [32] - a simple pinhole camera.

Uncertainty. As well as predicting per-pixel TOC values, our predictor also provides per-pixel TOC uncertainties, $\boldsymbol{\sigma_t}$. We can use this to weight our reprojection prediction. To do this, we need to propagate the uncertainty to calculate an uncertainty in pixel position, $\boldsymbol{\sigma}_{\hat{\boldsymbol{p}}}$. We use auto-gradient computation in PyTorch [30] to calculate the Jacobian $\mathbf{J}$ of the transformation from $\mathbf{t}$ to $\hat{\boldsymbol{p}}$. Next, we transform the uncertainty using the first order approximation given by Ochoa and Belongie [28],

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{p}}} \approx \mathbf{J}\boldsymbol{\Sigma}_{\mathbf{t}}\mathbf{J}^T, \quad \text{where } \boldsymbol{\Sigma}_{\boldsymbol{x}} = \operatorname{diag}(\boldsymbol{\sigma}_x^2). \tag{5}$$

Training loss. We now have a predicted pixel position $\hat{\boldsymbol{p}}$, and associated uncertainty $\boldsymbol{\sigma}_{\hat{\boldsymbol{p}}}$. We construct a loss which minimizes the average projected pixel error, weighting samples according to their uncertainty,

$$\mathcal{L}_{\text{FOCUS-O}} = \frac{1}{NP}\sum_{i=1}^{NP}\left\|\frac{\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i}{\boldsymbol{\sigma}_{\hat{\boldsymbol{p}}_i}}\right\|_2. \tag{6}$$

We minimize this loss by optimizing the parameters r, s, t, z_p, z_s. Similarly to FOUND [13], we use a two-stage optimization process - first optimizing just registration parameters r, s, t, and then all parameters. Each stage runs for 500 epochs, using an Adam optimizer [21] with a learning rate of 0.001.

4. Experiments

3D baseline. We evaluate on Foot3D, a baseline 3D foot reconstruction dataset released in [13] - 14 scenes of real scanned feet, with a total of 474 calibrated images. All methods are provided with the same set of input images and camera calibrations.

Method	NN chamfer error (mm) ↓			NN normal error (°) ↓		
	Mean	Median	RMSE	Mean	Median	RMSE
COLMAP	**1.8**	**0.9**	3.1	21.2	13.5	30.4
FOUND	2.6	2.2	3.3	**13.4**	**9.9**	19.5
FOCUS-SfM	2.0	1.5	**2.7**	14.1	11.1	**18.0**
FOCUS-O	2.1	1.8	**2.7**	13.5	10.2	18.8

Table 1. **3D reconstruction results.** Our methods yield the lowest RMSE chamfer error. While COLMAP still leads on mean and median chamfer error, it performs substantially worse on surface normal error, showing a substandard surface reconstruction. Our methods perform comparably to FOUND on surface normal reconstruction, despite FOCUS-SfM using no 3D prior in the Poisson reconstruction stage, and FOCUS-O not using surface normals during optimization.

As the task focuses on reconstruction of the foot (not the leg), we cut off all meshes at a height of 10 cm for evaluation. We evaluate the quality of our 3D optimization method by comparing the output mesh to a ground truth scan from our dataset. We select a sample of 10,000 points from each mesh, sampling uniformly over the surface area. For each point, we find the nearest neighbor (NN) on the surface of the other mesh, and capture the difference in Euclidean distance (chamfer error), and the angular difference between the surface normals at the two points (normal error).

We report the statistics of the 20,000 samples across the two meshes. We compare our methods against photogrammetry pipeline COLMAP [32, 33], and surface normal optimization method FOUND [13].

Speedtesting. We evaluate the speed and requirements of all methods. All methods were evaluated on the same machine, with an NVIDIA Quadro P2200 GPU.

Number of views. We vary the input views available to each method to investigate how this affects reconstruction quality. Where views are reduced, we sample views evenly in the left-right direction.

5. Results

Reconstruction accuracy. We show in Table 1 the accuracy of our two reconstruction methods, compared to FOUND and COLMAP, when all views are available (25-40) for each scan.

When all views are available, COLMAP performs the best for mean and median chamfer distance. However, our methods perform better on chamfer RMSE as COLMAP is prone to noise in reconstruction. Furthermore, our methods far outperform COLMAP with regards to surface normal quality, and in fact perform comparably with FOUND, which relies entirely on surface normals for fit-

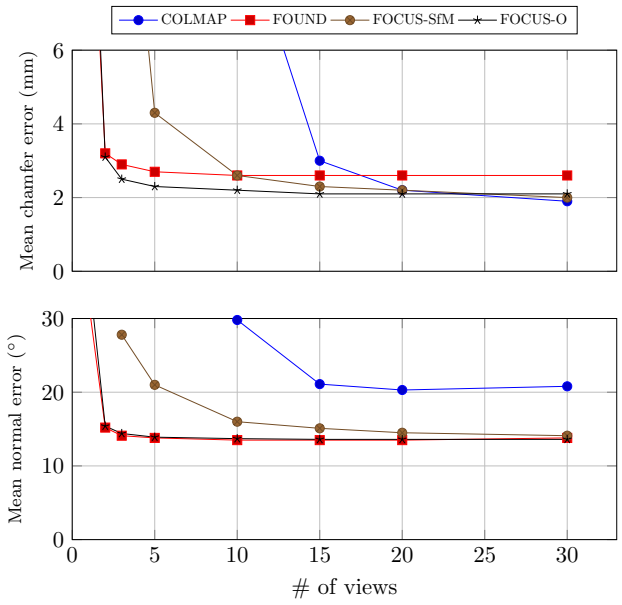

Figure 7. **Reconstruction quality as number of views varies.** FOCUS-O shows the best performance under a few-view setting, retaining its accuracy with as few as 3 views. FOCUS-SfM performs substantially better than COLMAP, showing accurate reconstruction performance with 10 views, compared to COLMAP requiring more than 15.

ting. FOCUS-O uses no surface normals in the reconstruction process - only requiring TOC correspondences.

Detailed qualitative results can be seen in Figure 8. These show that COLMAP fits often exhibit noise around edges, which is problematic for taking measurements. Our methods are also superior at capturing the surface around the toes - the separation of the toes is handled better by both FOCUS-SfM and FOCUS-O.

FOCUS-O tends to have fewer large sources of error both for normal and chamfer, providing a more reliable result. This approach, along with FOUND, produces watertight, parameterized meshes as output.

Number of views. We show the performance under a varying number of views in Figure 7. The strong 3D priors, and use of uncertainty, in FOUND and FOCUS-O ensure that they retain their reconstruction accuracy with as few as 3 views. COLMAP has the highest view requirement, showing a significant reduction in accuracy (and some outright failures in reconstruction) for fewer than 15 views. While reconstructing using some similar principles to COLMAP, FOCUS-SfM is able to take advantage of the rich information provided by TOCs to produce accurate reconstructions with as few as 10 views.

Method	Inference time (s) 10-view	Inference time (s) all-view	GPU?	Differentiable rendering?
COLMAP	-	465	✓	✗
FOUND	260	260	✓	✓
FOCUS-SfM	**22**	**89**	✗	✗
FOCUS-O	40	100	✓	✗

Table 2. **Speed and requirements.** FOCUS-O and FOCUS-SfM are substantially faster than existing methods. Furthermore, neither require differentiable rendering, and FOCUS-SfM does not require a GPU.

	NN mean chamfer error (mm) ↓	NN mean normal error (°) ↓
FOCUS-SfM	**2.0**	**14.0**
w/o subpixel matching	2.4	15.4
w/o normal aggregation	2.5	107.5

Table 3. **FOCUS-SfM ablation study.** Both the subpixel matching and normal aggregation steps of FOCUS-SfM are crucial for its quantitative reconstruction performance. The drastic increase in normal error without normal aggregation is partly due to the reconstruction algorithm occasionally inverting the surface normals.

	NN Chamfer error (mm) ↓		NN Normal error (°) ↓	
	3 view	20 view	3 view	20 view
FOCUS-O	**2.5**	**2.1**	**14.4**	**13.5**
w/o uncertainty	2.7	2.2	15.0	13.9

Table 4. **FOCUS-O ablation study.** The use of TOC uncertainty in our optimization process improves all reconstruction metrics, both for a low and high view count.

Implementation. Table 2 compares the inference times and computational requirements of all methods. Both of our methods are substantially faster than COLMAP and FOUND. Compared to FOUND, neither require differentiable rendering, drastically reducing the memory requirements, and making them easier to implement on devices that would not support it. Further, FOCUS-SfM is a CPU based method (except for the initial TOC predictions), so is possible to run on devices without GPU optimization support, including mobile devices.

Ablation study. We show results of our ablation for both FOCUS-SfM and FOCUS-O in Tables 3 and 4 respectively.

For FOCUS-SfM, aggregating the predicted surface normals is critical to capturing geometry in certain areas of the foot, especially around the toes, as naive Poisson reconstruction is not capable of estimating that detail. Furthermore, Poisson reconstruction may generate a plausible mesh with inverted face normals, hence the high surface normal error. Our subpixel matching is also critical for cap-

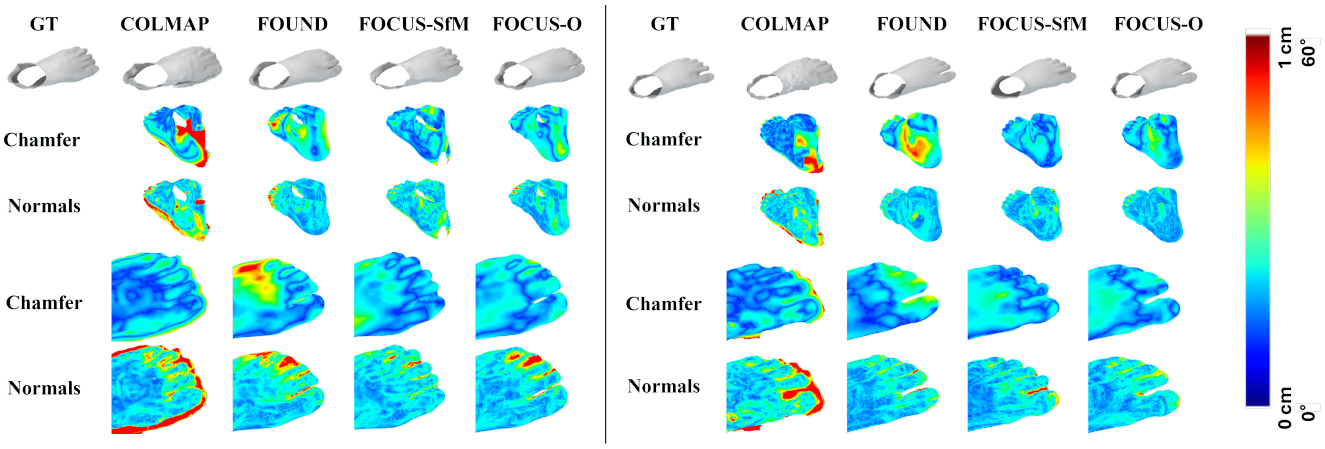

Figure 8. **Qualitative reconstruction results.** The reconstruction quality is compared across two scans in the Foot3D dataset, comparing COLMAP, FOUND, FOCUS-SfM and FOCUS-O. COLMAP is prone to noise around the foot boundaries, and captures less surface detail in particular around the toes and heel. Further qualitative comparisons can be found in the supplementary material.

turing some of the finer details on the foot surface.

For FOCUS-O, the use of uncertainty in weighting the various TOC samples used for fitting the FIND model provides a boost to fitting accuracy.

6. Conclusion

We have shown that the TOC representation provides a strong signal for predictors to identify dense correspondences on in-the-wild images, and that it is possible to train these models using entirely synthetic data. Our synthetic dataset, SynFoot2, provides substantial improvements on SynFoot, including more background and lighting variation, and adding articulation to the 3D models used for rendering.

We have shown that TOCs can be combined with uncertainty and surface normal predictions to recover high quality foot reconstructions from multiple views.

The first of our reconstruction methods, FOCUS-SfM, effectively identifies subpixel TOC matches across multiple images, and triangulates these to provide a high quality reconstruction with no enforcement of a 3D prior on the output foot shape. This method requires fewer views than COLMAP, and can operate entirely on a CPU.

Our second method, FOCUS-O, is capable of directly optimizing the parameterized FIND model to match the TOC predictions across images. FOCUS-O effectively uses TOC uncertainty, and provides a superior reconstruction to FOUND, while matching the low view requirement, and running substantially faster.

Each method has its own advantages. FOCUS-SfM may be suitable for applications where speed is desired, no GPU is available, and only the visible portion of the surface is of interest to downstream tasks. In contrast, FOCUS-O may be more desirable when a fully parameterized and watertight mesh is required.

7. Limitations and future work

Our methods currently rely on calibrated cameras for reconstruction. However, the dense correspondences could be used to recover the relative camera poses, up to a scale ambiguity. We did not explore this as part of this paper, as our evaluations are against Foot3D, requiring accurate camera scale and pose. This approach is implemented in the code released alongside this paper, and can be run on an arbitrary set of uncalibrated images.

FOCUS-O does not leverage the image surface normal predictions, as there is no simple mapping from TOC to surface normals, and so would require differentiable rendering. Modifying the approach to use surface normals could improve accuracy, at the cost of computation time. Similarly, FOCUS-SfM does not use the predicted TOC uncertainty, as we did not find it to improve the reconstruction. Future research could identify an approach to use this signal to improve FOCUS-SfM's reconstruction.

8. Acknowledgments

The authors acknowledge the collaboration and financial support of Trya Srl.

References

[1] Artec Leo, Artec3D. https://www.artec3d.com/portable-3d-scanners/artec-leo. Accessed: 2024-08-01. 1

[2] Volumental In-store Scanner, Volumental. https://volumental.com/volumental-instore. Accessed: 2024-08-01. 1

[3] Poly Haven. https://polyhaven.com. Accessed: 2024-08-01. 4

[4] Xesto. https://www.xesto.io. Accessed: 2024-08-01. 1, 3

[5] I Alhashim. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 4

[6] 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, ICVS 2008 Santorini, Greece, May 12-15, 2008 Proceedings 6*, pages 161–170. Springer, 2008. 3

[7] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021. 4

[8] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 2

[9] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007. 3

[10] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D.S. Seljebotn, and K. Smith. Cython: The best of both worlds. *Computing in Science Engineering*, 13(2):31 –39, 2011. 5

[11] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975. 5

[12] Oliver Boyne, James Charles, and Roberto Cipolla. FIND: An unsupervised implicit 3D model of articulated human feet. In *British Machine Vision Conference (BMVC)*, 2022. 1, 2, 3, 4, 6

[13] 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 (WACV)*, 2024. 1, 2, 3, 4, 5, 6

[14] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*. The Eurographics Association, 2008. 5

[15] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[17] Can Gümeli, Angela Dai, and Matthias Nießner. Objectmatch: Robust registration using canonical object correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13082–13091, 2023. 2

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[19] Peggy A Houglum and Dolores B Bertoti. *Brunnstrom's clinical kinesiology*. FA Davis, 2011. 3, 4

[20] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2, 5

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

[22] 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, 2, 3

[23] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1649–1652, 2009. 5

[24] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[26] 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*, pages 2300–2305, 2017. 1, 3

[27] 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

[28] Benjamin Ochoa and Serge Belongie. Covariance propagation for guided matching. In *Proceedings of the Workshop on Statistical Methods in Multi-Image and Video Processing (SMVP)*, 2006. 6

[29] 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*, pages 568–585. Springer, 2022. 3

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[31] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2

[32] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 6

[33] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 2, 6

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[35] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016. 2

[36] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–110. IEEE, 2012. 3

[37] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 2

[38] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[39] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3756–3764, 2015. 2

[40] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2

[41] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[42] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7054–7063, 2020. 3